UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JULLYANA MATTOS VYCAS

ANÁLISE DO DESEMPENHO ACADÊMICO NA CIÊNCIA DA COMPUTAÇÃO - UFRJ

RIO DE JANEIRO

2018

JULLYANA MATTOS VYCAS

ANÁLISE DO DESEMPENHO ACADÊMICO NA CIÊNCIA DA COMPUTAÇÃO - UFRJ

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. João Carlos Pereira da Silva, D.Sc.

RIO DE JANEIRO

2018

CIP - Catalogação na Publicação

Vycas, Jullyana Mattos
V996a Análise do Desempenho Acadêmico na Ciência da Computação - UFRJ / Jullyana Mattos Vycas. -- Rio de Janeiro, 2018.
64 f.

Orientador: João Carlos Pereira da Silva.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2018.

1. Data Warehousing. 2. Business Intelligence. 3. Data profiling. I. Pereira da Silva, João Carlos, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a).

JULLYANA MATTOS VYCAS

ANÁLISE DO DESEMPENHO ACADÊMICO NA CIÊNCIA DA COMPUTAÇÃO - UFRJ

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em _____ de _____ de _____ .

BANCA EXAMINADORA:

_____

Prof. João Carlos Pereira da Silva, D.Sc.

_____

Prof. Maria Helena Cautiero Horta Jardim, D.Sc.

_____

Prof. Maria Luiza Machado Campos, D.Sc.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer ao meu extraordinário orientador, que desde o início se mostrou compreensivo e prestativo. Essa jornada teria sido muito mais difícil sem a sua excelente mentoria.

Agradeço também ao SIGA, por fornecer os dados que tornaram esse projeto possível, à professora Maria Luiza, por me ajudar com a modelagem dimensional logo no início projeto e ao meu orientador acadêmico, o professor Claudson Bornstein, sem o qual não teria tido uma experiência tão positiva na graduação.

Por fim, gostaria de agradecer aos meus pais, por terem me apoiado por toda a minha vida, e ao meu marido, por ter preservado a minha sanidade nos momentos mais penosos.

RESUMO

Traçar o perfil do corpo estudantil, assim como seu quadro geral, é indispensável para a identificação de seus problemas. Professores tem acesso a muitos dados dos alunos, mas poucas ferramentas de análise estão disponíveis. Este trabalho propõe uma modelagem dimensional que estrutura os dados existentes de forma a facilitar a análise do desempenho acadêmico na Ciência da Computação. Uma versão inicial de banco analítico foi criada e populada para este estudo, disposta juntamente de um conjunto básico de análises e seus respectivos resultados.

Palavras-chave: Data Warehousing. Business Intelligence. Data profiling.

ABSTRACT

Tracing the profile of the student body and its general situation is indispensable to identify its problems. Instructors have access to a lot of data, but not many analyzing tools are available. This work proposes a dimensional model to structure existing data in a way to facilitate the analysis of the performance of the Computer Science students. An initial version of the analytical database was created and populated for this study, along with a set of analysis and the respective results.

Keywords: Data Warehousing. Business Intelligence. Data profiling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CR	Coefieciente de Rendimento
DW	Data Warehouse
ETL	Extract Transform Load
SIGA	Sistema Integrado de Gestão Acadêmica
SiSU	Sistema de Seleção Unificada
SQL	Structured Query Language

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Uma parcela significativa de ingressantes no ensino superior nunca recebe seu diploma. Em cursos relacionados a Computação, a evasão é maior do que a média nacional, tanto em instituições privadas quanto em instituições públicas [5]. Métricas semelhantes podem ser observadas dentro da UFRJ, onde cerca de metade dos alunos não concluem o curso de Ciência da Computação.

Para tratar o problema, é necessário buscar suas raízes. Em menor escala, é possível observar desistências dentro do curso. Trancamentos e abandonos em disciplinas podem ser sintomas iniciais da evasão do ensino superior como um todo. Em grande escala, encontramos pontos críticos na graduação verificando a quantidade de alunos inscritos através do tempo.

Vivemos em um mundo repleto de dados, mas com pouca informação. Esta é a realidade dos docentes do curso de Ciência da Computação na UFRJ, que dispõe de dados de centenas de discentes, porém carecem de informação sobre o estado geral do curso. Atualmente, orientadores e coordenadores acadêmicos dispõe de poucas ferramentas para ajudá-los a compreender a visão geral da curso. Foi desta necessidade que surgiu a ideia de modelar, estruturar e analisar o quadro atual da graduação.

1.2 OBJETIVOS

Este trabalho busca retratar a situação do curso de Ciência da Computação na UFRJ, estudando o desempenho dos alunos ao seu decorrer e os principais pontos de desistência. Com este retrato, podemos fazer comparações entre o que é esperado do aluno e a realidade encontrada.

O modelo proposto neste trabalho pode servir no futuro como mecanismo para auxiliar professores a acompanhar o rendimento dos alunos, verificar repercussões de

decisões da coordenação e até mesmo descobrir novos aspectos da vida acadêmica.

Informações tais como a taxa de evasão por período, desempenho acadêmico e tempo médio de conclusão de curso são importantes para traçar o perfil do aluno de graduação de Ciência da Computação, entender a situação geral do corpo discente e tomar decisões que aumentem a qualidade do curso.

1.3 ORGANIZAÇÃO

Separamos este trabalho em três capítulos principais. No Capítulo 2, descrevemos os dados que nos foram disponibilizados e o tratamento que lhes foi dado, bem como a transição de dados operacionais para um banco de dados puramente analítico. Em seguida, no Capítulo 3, propomos uma modelagem dimensional e descrevemos como ela foi implementada. Finalmente, o Capítulo 4 mostra análises feitas com base no banco desenvolvido, resultados encontrados e especulações que podem ser feitas.

2 DADOS RECEBIDOS E TRATAMENTO

Para a realização do nosso estudo, preparamos um banco de dados analítico. Neste capítulo, descreveremos os dados obtidos, seu estado original e o tratamento que lhes foi aplicado.

2.1 DADOS OBTIDOS

Os dados utilizados neste trabalho foram extraídos de um banco de dados operacional do SIGA. Pre-processados, os dados foram anonimizados antes de serem fornecidos ao projeto, de forma a proteger a identidade dos alunos envolvidos. Nenhum dado pessoal identificável foi disponibilizado.

Doze planilhas foram disponibilizadas para este estudo, todas em formato Excel (xls e xlsx). No apêndice 5.3, apresentamos exemplos de cada planilha de dados recebida.

No total, recebemos dados de 2012 alunos, relativos aos anos de 2000 até 2017, incluindo:

- Carga Horária Obtida Acumulada
- CR Acumulado por Período
- CR do Período
- Créditos Obtidos Acumulados
- Disciplinas Cursadas com Nota e Situação Final
- Forma de Ingresso
- Histórico das Disciplinas
- Matérias Trancadas pelos Alunos
- Períodos com CR Menor que 3

- Períodos de Cancelamento por Abandono
- Períodos de Trancamento de Matrícula

Algumas informações interessantes não estavam disponíveis em nosso conjunto inicial. Nenhuma das tabelas recebidas do SIGA dizia quantos créditos uma disciplina está associada a, por exemplo, ou quais disciplinas são obrigatórias para o Bacharelado em Ciência da Computação. Contudo, todo dado que faltava já havia sido disponibilizado publicamente pelo SIGA, de fácil acesso. Bastou, portanto, extrair esses dados para arquivos de texto e planilhas que posteriormente serviram de entrada para a ferramenta de ETL escolhida. Os seguintes dados foram originados desse processo:

- Departamentos e Códigos Associados
- Carga Horária e Créditos de Cada Disciplina
- Lista de Disciplinas Obrigatórias (Ciência da Computação)

Todas as fontes de dados utilizadas neste trabalho são não-relacionais, ou seja, não foram extraídas diretamente de um banco relacional, mas sim encontradas em planilhas ou arquivos de texto. Felizmente, a falta de um banco relacional não introduz desvantagens relevantes. De fato, o poder das ferramentas de ETL ao lidar com fontes heterogêneas de dados minimiza a necessidade de guardar todos os dados em um banco único [6].

2.2 TRATAMENTO DOS DADOS

2.2.1 Extração, Transformação e Carregamento

Dados operacionais são a fonte principal para bancos analíticos. Contudo, o processo de criação de um data warehouse não constitui somente de uma cópia direta do banco operacional. De fato, a disparidade entre o operacional e o analítico é tão grande que podemos dizer que são mundos diferentes.

No mundo operacional, a manipulação é feita linha por linha. Consultas são feitas um registro de cada vez, parte de um fluxo de transações sob restrições de seu sistema operacional [7]. As prioridades são perfomance e disponibilidade. Os dados estão em seu estado mais cru, possivelmente sem descrições cuidadosas ou legendas explanatórias.

O mundo analítico, por sua vez, preza pela fácil compreensão dos dados. Seu conteúdo deve ser intuitivo e óbvio não só para o desenvolvedor, mas também para o usuário [7]. Redundâncias não são um problema para bancos analíticos; de fato, redundâncias são uma vantagem, tornando consultas mais eficientes (menos junções de tabelas) e esclarecendo o significado dos dados.

O processo de extração, transformação e carregamento ("Extract-Transformation-Load" ou ETL) é a ponte entre os dois mundos, representada pela Figura 1. Uma analogia interessante feita pelo Kimball Group [7] é de que o estágio de ETL se assemelha à cozinha de um restaurante. Nele, dados crus vindos do banco operacional são preparados e transformados em refeições. Inconsistências são tratadas, diferentes fontes são combinadas e o resultado final deve ser facilmente compreendido por alguém que não trabalhe diretamente com o banco.

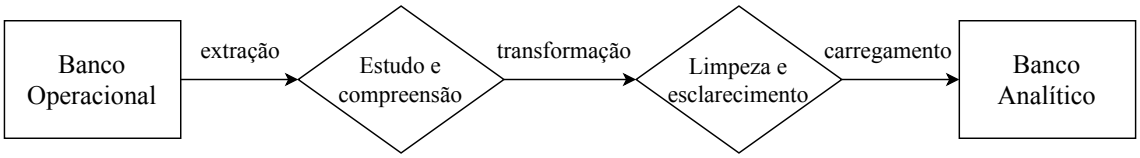


Figura 1: Extração, transformação e carregamento.

A extração dos dados começa pela leitura e compreensão dos dados operacionais. Tudo aquilo que for considerado relevante para o data warehouse será copiado e levado para o próximo estágio: a transformação. É nela que acontece a limpeza (correção de erros de digitação, formatação padrão, resolução de entradas nulas ou desaparecidas), a combinação das diversas fontes, remoção de dados repetidos e criação de chaves para o banco analítico [7]. Só então os dados serão carregados nos data marts.

2.2.2 Data Profiling

Antes de fazer qualquer tipo de tratamento em cima dos dados disponíveis, buscando aumentar sua qualidade e garantir sua consistência, é necessário estudar seu estado inicial em detalhes. É interessante observar quais colunas das planilhas podem ser nulas, que tipo de dependência existe entre diferentes planilhas, qual a faixa de valores relacionada a cada métrica, entre outros detalhes.

Evidentemente, é possível extrair todas essas informações através de consultas aos dados e observação manual. De fato, algumas anomalias foram encontradas desta forma. Contudo, para obter um resultado melhor embasado, é mais adequado utilizar uma ferramenta de data profiling.

Data profiling é o processo de coleta de estatísticas e outras informações sobre os dados disponíveis em diferentes fontes [2]. Em outras palavras, é o processo de conhecimento dos dados, a partir do qual podemos obter mais dados (agregação, média, número de valores distintos, faixa de valores) e compreender a natureza dos dados existentes.

Vamos considerar um exemplo encontrado durante a observação inicial dos dados. Na tabela *Disciplinas Cursadas em Cada Período*, algumas entradas estão associadas a um período 0. Na UFRJ, existem três possíveis períodos: 1 (primeiro semestre), 2 (segundo semestre), 3 (recesso). Surge o questionamento: o que significa período 0?

Quando um aluno pede transferência de créditos, tanto de disciplinas cursadas em outras instituições quanto dentro da UFRJ, a equivalência é lançada em seu boletim associada a um ano, porém sem um período específico. A hipótese era de que esses casos correspondiam às entradas com período 0 na tabela de disciplinas cursadas.

De fato, das 2.694 entradas associadas ao período 0, 2.344 (87%) possuem conceito T, que remete a "transferido". Infelizmente, os 350 (13%) casos restantes tinham conceitos entre 0 e 100. Ao repararmos nestes casos anômalos, percebemos que todos estavam associados a um número pequeno de alunos (39). Com isso,

podemos inferir que estas entradas se referem a alguns poucos casos especiais de transferências internas cujo conceito foi mantido. Para assegurar tratamento similar a todas as transferências, estabelecemos que todas as entradas com período 0 teriam conceito T.

O exemplo anterior mostra como o estudo dos dados pode ser feito manualmente. Este processo se mostra tedioso, ineficiente em muitos casos e ineficaz. Entra, então, a ferramenta de data profiling. Escolhemos DataCleaner, da comunidade eobjects.org, indicada por [2].

2.2.3 Estudo dos Dados

Cada uma das tabelas recebidas foi cuidadosamente estudada utilizando a ferramenta DataCleaner. Vamos utilizar uma tabela simples (Carga Horária Acumulada) como exemplo do processo, representado na figura 2. Esta tabela tem apenas dois campos: 'aluno' e 'carga_horaria', e precisamos determinar quais verificações precisam ser feitas sobre esses dados.

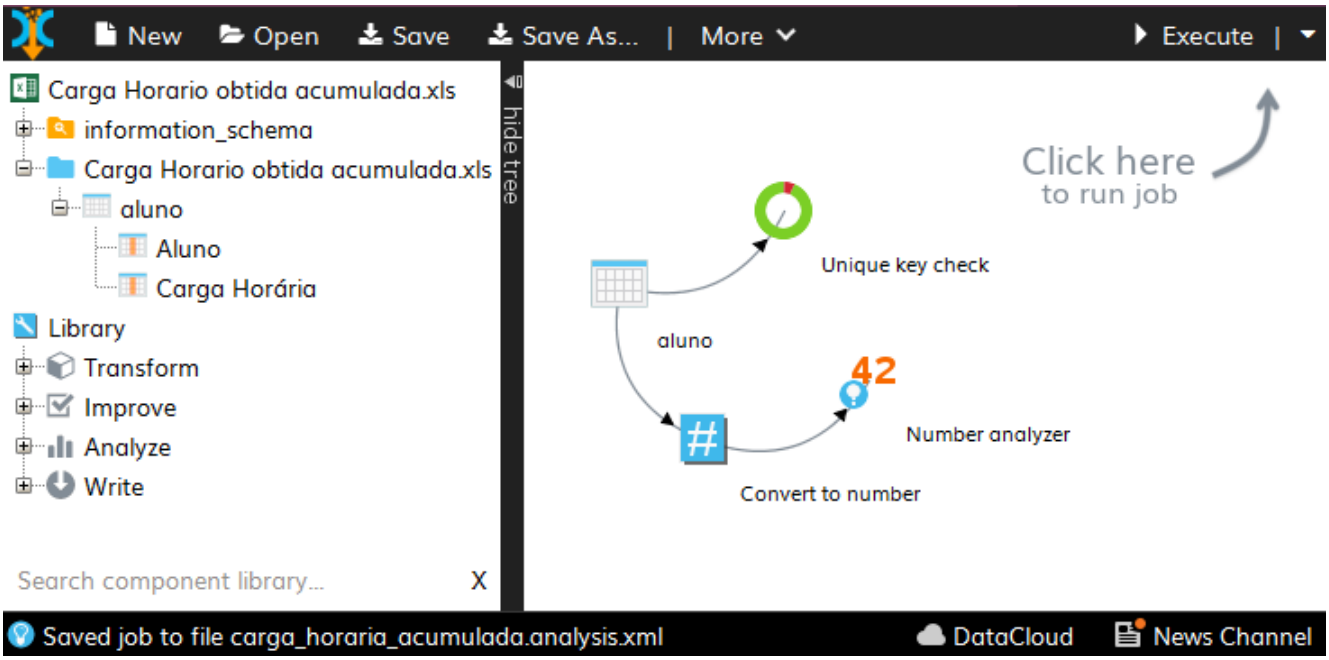
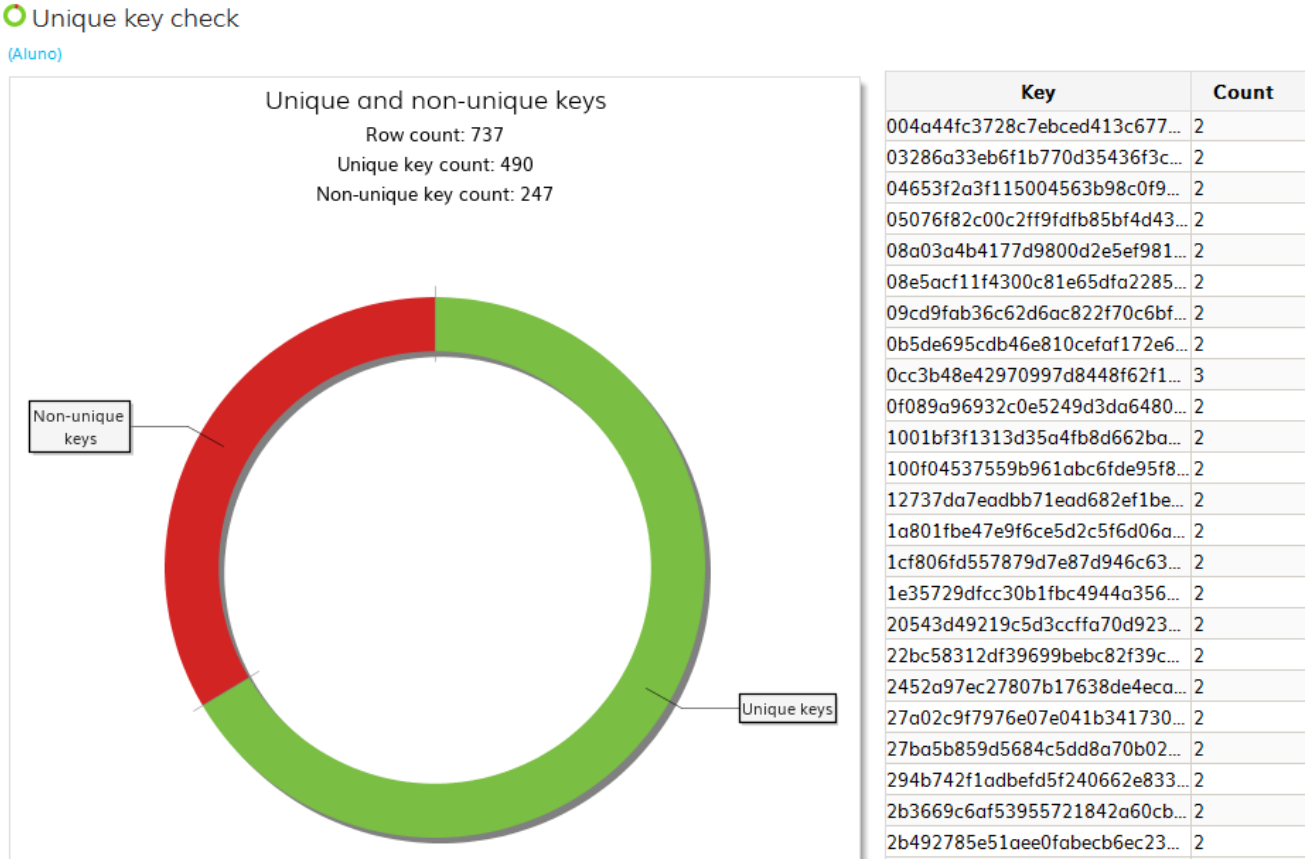


Figura 2: Exemplo de uso da ferramenta DataCleaner

No caso do campo 'aluno', precisamos saber se existem múltiplas entradas asso-

ciadas a um mesmo aluno. Para isso, adicionamos uma verificação de chave única ("Unique key check") na ferramenta. O resultado, apresentado na figura 3, mostra que 247 das 737 entradas não possuem chave única, ou seja, alguns alunos aparecem duas vezes nessa tabela. Inspecionando esses casos, percebemos que alguns são meras duplicatas (linhas iguais na tabela).



Key	Count
004a44fc3728c7ebced413c677...	2
03286a33eb6f1b770d35436f3c...	2
04653f2a3f115004563b98c0f9...	2
05076f82c00c2ff9fdfb85bf4d43...	2
08a03a4b4177d9800d2e5ef981...	2
08e5acf11f4300c81e65dfa2285...	2
09cd9fab36c62d6ac822f70c6bf...	2
0b5de695cdb46e810cefaf172e6...	2
0cc3b48e42970997d8448f62f1...	3
0f089a96932c0e5249d3da6480...	2
1001bf3f1313d35a4fb8d662ba...	2
100f04537559b961abc6fde95f8...	2
12737da7eadbb71ead682ef1be...	2
1a801fbe47e9f6ce5d2c5f6d06a...	2
1cf806fd557879d7e87d946c63...	2
1e35729dfcc30b1fbc4944a356...	2
20543d49219c5d3ccffa70d923...	2
22bc58312df39699bebc82f39c...	2
2452a97ec27807b17638de4eca...	2
27a02c9f7976e07e041b341730...	2
27ba5b859d5684c5dd8a70b02...	2
294b742f1adbefd5f240662e833...	2
2b3669c6af53955721842a60cb...	2
2b492785e51aee0fabecb6ec23...	2

Figura 3: Resultado da verificação de chave única (unique key check).

Uma vez detectado o problema, devemos pensar em uma forma de tratamento. Imaginamos que as linhas extras para alguns alunos são atualizações em sua carga horária acumulada, e que os menores valores são um histórico. Infelizmente, não há um período associado a esses valores. Portanto, não é interessante guardar esse histórico, e decidimos manter apenas o maior valor para cada aluno.

Já no caso do campo 'carga_horaria', gostaríamos de fazer uma verificação geral de números. Por exemplo, queremos determinar se existem valores nulos e quais são o maior e menor valores encontrados. Precisamos primeiramente converter o campo atual (texto) em numeração ("Convert to number"). Em seguida, adicionamos um

analisador de números (“Number analyzer”). A ferramenta nos dá um relatório completo, apresentado na figura 4, dizendo qual o mínimo encontrado (0), o máximo encontrado (3.390), o número de entradas nulas (0) entre outras informações interessantes, como a média (1.451,961) e o desvio padrão (987,135). Esses dados serão úteis na hora de dividir esses valores numéricos em faixas adequadas para análise.

Number analyzer

(Carga Horária (as number))

	Carga Horária (as number)
Row count	737
Null count	0
Highest value	3.390
Lowest value	0
Sum	1.070.095
Mean	1.451,961
Geometric mean	0
Standard deviation	987,135
Variance	974.435,247
Second moment	717.184.341,859
Sum of squares	2.270.920.175

Figura 4: Resultado do analisador de números (number analyzer).

2.2.4 Transformações realizadas

Após inspeção de todos os dados recebidos, definimos que as seguintes transformações deveriam ser realizadas para tornar os dados consistentes e facilmente compreensíveis:

- Remover duplicatas de todas as tabelas;
- Considerar apenas a maior carga horária acumulada por aluno;
- Conceito em período 0 deve ser nulo;

- Período 0 deve se tornar “Transferido”;
- Caso existam dois ou mais registros de uma disciplina cursada por um aluno no mesmo perído, manter apenas o registro com maior conceito [1];
- Nomes de professores, inicialmente em caixa alta, serão capitalizados;
- Se uma disciplina é especial (Projeto Final, Monitoria) e não está associada a um dia da semana, substituir nulo por “Não Aplicável”.

O volume de dados durante o processo de ETL costuma ser o maior desafio a ser enfrentado. Em geral, a solução para este problema é pré-processar os dados, guardando cópias de segurança antes de transformações importantes, servindo como pontos de recuperação para caso alguma falha ocorra durante o carregamento [6].

Devido ao escopo do trabalho, porém, a quantidade de dados observados foi relativamente pequena comparada a grandes projetos empresariais de data warehousing. Por esse motivo, o tempo médio de execução das transformações necessárias e atualizações ao banco se manteve na ordem de minutos. Falhas não são dispendiosas pois recomeçar do início é barato em termos de processamento. Nesse caso, o passo de pré-processamento pode ser completamente ignorado sem consequências negativas.

Com isso, concluímos o tratamento dos dados recebidos. No próximo capítulo, descreveremos a modelagem e a criação do banco analítico.

[1] Neste caso, é provável que as notas não tenham sido lançadas no período correto, de forma que o primeiro lançamento (em geral, com conceito zero) tenha sido apenas provisório.

3 MODELAGEM E IMPLEMENTAÇÃO

Para a modelagem do banco de dados analítico, escolhemos seguir a modelagem dimensional de Ralph Kimball [8], amplamente adotada na área. Apesar de Kimball não ter elaborado os conceitos de dimensão e fato, não há dúvidas de que a metodologia do autor seja a mais renomada atualmente.

3.1 ELEMENTOS PRINCIPAIS

A modelagem dimensional tem dois elementos principais: as tabelas fato e as tabelas dimensão. A tabela fato é a tabela primária de um modelo dimencional, onde métricas de performance numéricas são armazenadas [7]. O fato, é, portanto, uma métrica do negócio representado. Como exemplo, em uma tabela de vendas diárias, uma métrica possível seria "quantidade vendida".

Tabelas fato estão conectadas a dimensões através de chaves estrangeiras. As dimensões dão detalhes sobre a situação na qual a métrica foi obtida. Voltando ao exemplo das vendas diárias, possíveis dimensões seriam dia, local, vendedor e produto. É através das dimensões que acessamos as tabelas fato. Afinal, ao fazer perguntas a um banco analítico, estamos especificando condições que queremos estudar. "Em qual dia da semana vendemos mais?", "Qual produto teve a maior queda em vendas?".

Podemos dizer, então, que tabelas dimensão são pontos de entrada para as tabelas fato. Dimensões são a interface de usuário para o data warehouse [7]. Os atributos de uma dimensão devem ser claros e auto-explicativos, sem abreviações ou códigos indecifráveis por si só.

Antes de apresentar o modelo dimensional proposto, é necessário descrever algumas particularidades na implementação de dimensões:

1. *Chave Substituta.* Toda dimensão deve ter uma chave primária única. Essa chave não pode ser a chave natural do sistema operacional, pois podem existir múltiplas

PK_PROFESSOR	NOME
1	Maria
2	João

Tabela 1: Professores

PK_PROFESSORES	FK_PROFESSOR
1	1
1	2

Tabela 2: Ponte Professores

linhas da dimensão associadas à mesma chave natural [8]. Por isso, criamos *chaves substitutas*, sobre as quais temos total controle. São simples inteiros atribuídos sequencialmente a cada entrada na tabela, começando com o valor 1.

2. *Dimensões com múltiplos valores.* Certos fatos estão associados a mais de um valor em uma mesma dimensão. Por exemplo, dois professores podem compartilhar uma turma. Contudo, a tabela fato está associada a apenas uma chave estrangeira referente a professor. Adicionar mais chaves estrangeiras é uma possível solução, mas a quantidade de professores não é constante. Portanto, precisamos de uma solução dinâmica.

   Nesses casos, a dimensão com múltiplos valores deve ser anexada à tabela fato através de uma chave de grupo [8]. Chamamos de ponte a tabela que guarda quais são esses grupos.

   Por exemplo, na Tabela 1 temos uma lista de professores. Se Maria e João foram responsáveis pela mesma turma em um período, a Ponte Professores seria preenchida como indicado na Tabela 2. Note que a chave primária de cada professor do grupo estará presente em uma linha da tabela ponte como chave estrangeira.

   Com isso, indicamos que o grupo com chave primária 1 é formado pelos professores Maria e João. Assim, basta apontar para essa chave na hora de associar a situação

em disciplina a um conjunto de professores e o problema foi resolvido.

3. *Faixas.* Ocasionalmente é interessante filtrar os resultados de uma consulta baseado em métricas agregadas. Por exemplo, podemos estar interessados em encontrar a disciplina com maior proporção de notas abaixo de 3. Para isso, precisamos agregar o fato conceito em faixas e disponibilizá-las como entradas para a nossa consulta, em uma dimensão [8].

3.2 MODELO DESENVOLVIDO

Dois fatores principais foram considerados durante a modelagem: os dados disponíveis e as perguntas que queremos responder através da análise dos mesmos. De forma geral, todo atributo que gostaríamos de utilizar como entrada para nossas análises deve estar associado a uma dimensão, e toda métrica que queremos observar deve ser um fato dentro de uma tabela fato. A Figura 5 mostra a modelagem estrela desenvolvida.

O apêndice 5.3 contém tabelas descritivas de cada dimensão e fato utilizados. São estes:

- Dimensão Aluno
- Dimensão Dia da Semana
- Dimensão Disciplina
- Dimensão Matrícula
- Dimensão Período
- Dimensão Professor
- Dimensão Situação Final
- Fato Situação em Disciplina
- Fato Situação em Período

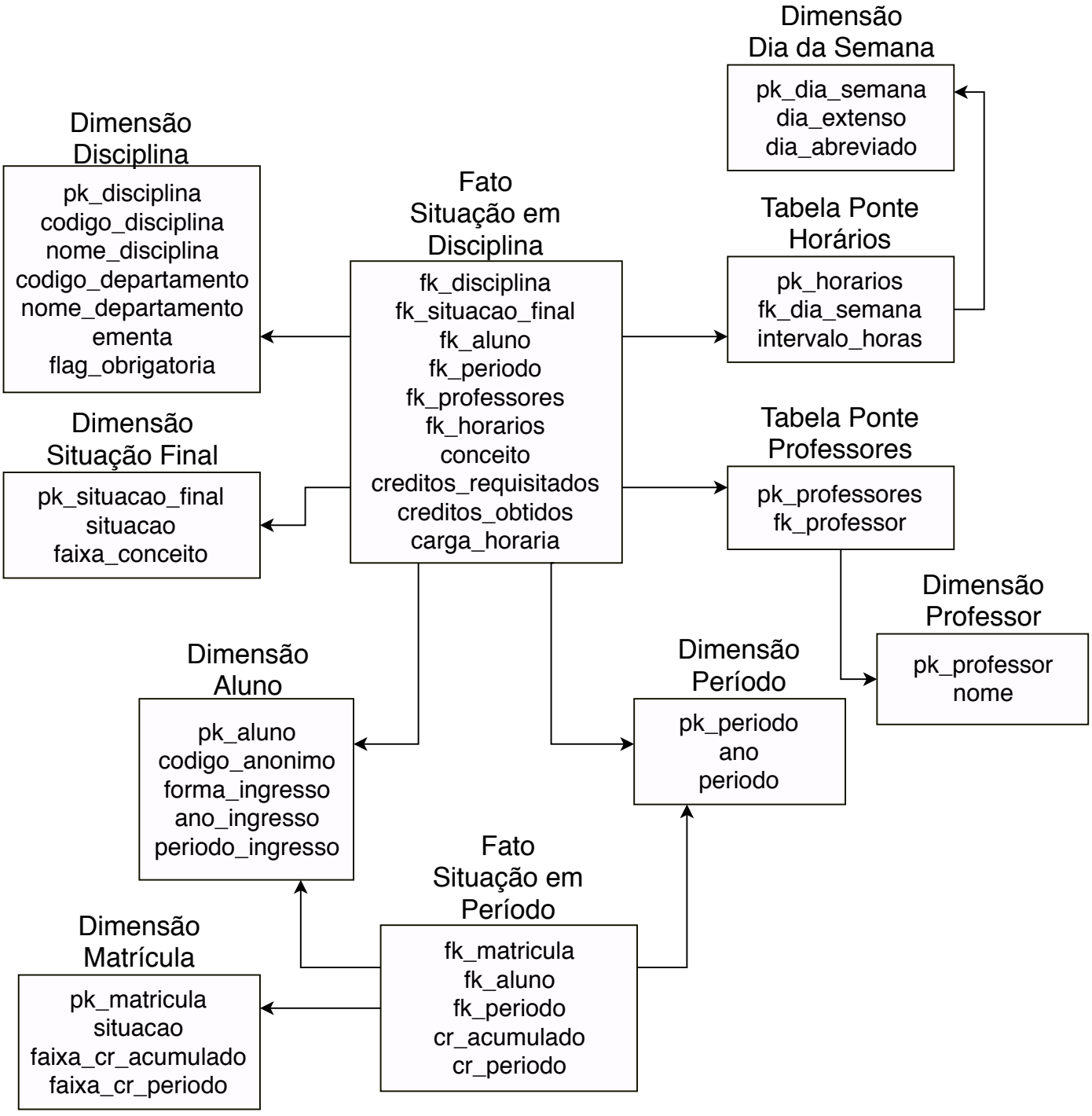


Figura 5: Modelagem Dimensional

Um detalhe relevante sobre a abordagem escolhida é o posicionamento dos valores carga horária e créditos. Apesar de ambos serem inerentes à disciplina cursada, por serem valores numéricos, são métricas de interesse para análise. Por esse motivo, eles foram dispostos na tabela fato.

3.3 IMPLEMENTAÇÃO DO MODELO

Para transformar a modelagem conceitual em um banco efetivo, utilizamos a ferramenta Power Architect [1]. Esta ferramenta, recomendada por [2], permite criar esquemas estrela. Uma vez que a modelagem esteja devidamente representada, é possível gerar um script em SQL que cria a estrutura do banco necessária.

A Figura 6 mostra a modelagem dimensional traduzida para a ferramenta.

Cada tabela ponte foi implementada usando duas tabelas no banco: dimensão grupo e ponte. Essa separação foi necessária pois na prática a chave primária de uma tabela ponte é composta pela chave do grupo e pela chave do indivíduo que faz parte do grupo. De forma a apontar apenas para a chave de grupo a partir de um fato, precisamos de uma tabela que contém apenas as chaves de grupo. Portanto, temos essa dimensão extra, que chamamos de dimensão grupo.

Com exceção das pontes, o mapeamento da modelagem foi direto. Foi necessário definir tipos para cada um dos atributos das tabelas. Essa definição costuma ser imediata: cadeias de texto são VARCHAR, números inteiros são INT, números com casas decimais são DOUBLE e flags são BOOLEAN.

Bons fatos devem ser numéricos, possibilitando agregações como soma, média, máximo e mínimo. Em transferências de créditos, o valor original do conceito era textual, “T”. Se mantivéssemos conceito como texto, estaríamos ferindo uma das recomendações principais de Kimball: métricas textuais devem estar em dimensões [7]. Além disso, perderíamos a capacidade de agregar conceitos. Dessa forma, conceito foi definido como DOUBLE, com transferências de valor nulo, indicando que não há conceito a ser analisado.

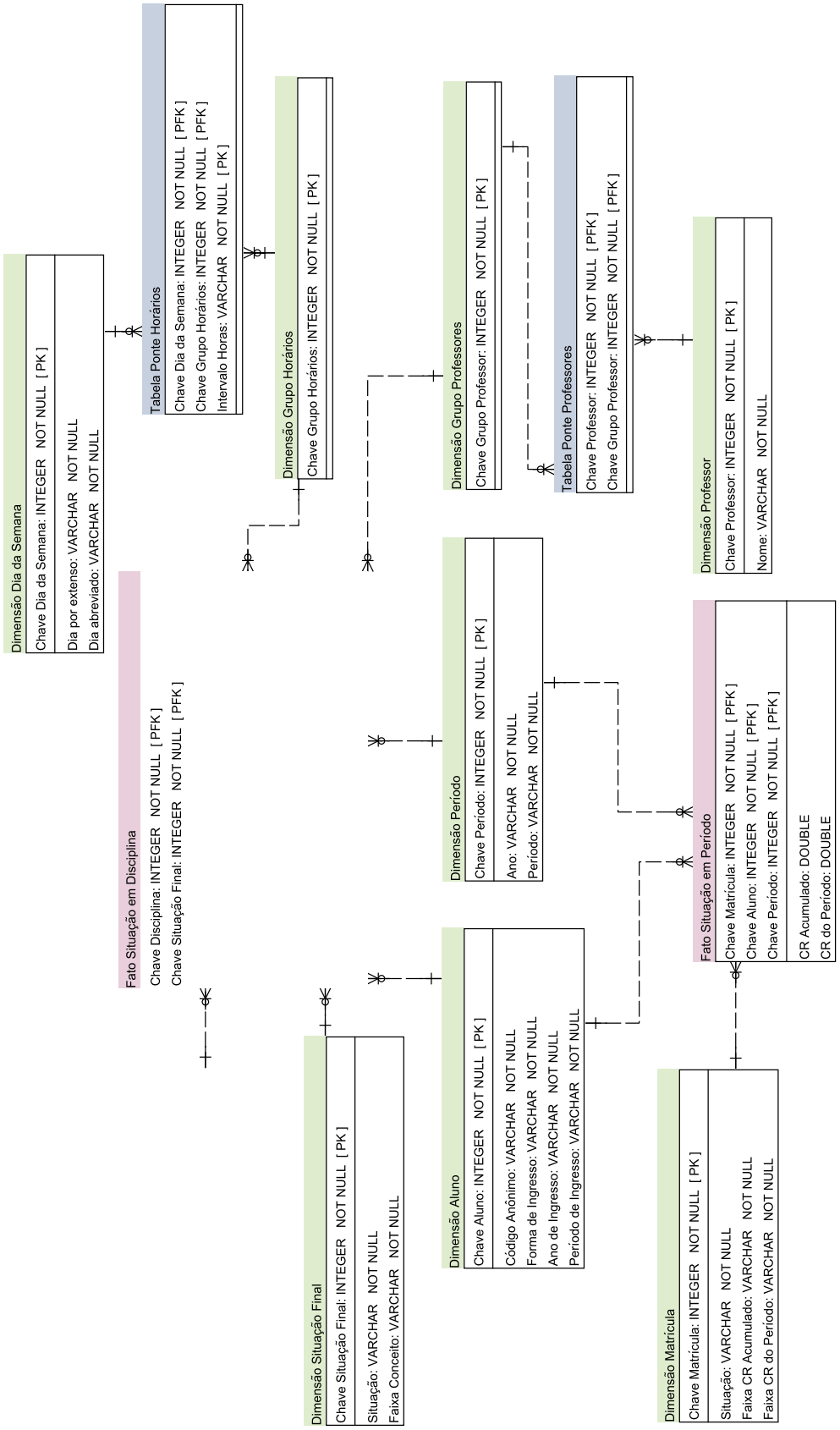


Figura 6: Modelagem representada na ferramenta SQL Power Architect.

3.4 PREENCHIMENTO DO BANCO DE DADOS

Para abrigar nosso banco de dados, escolhemos o sistema PostgreSQL [9]. Além de open source, esse sistema é facilmente integrável com a ferramenta que utilizaremos para popular o banco e fazer análises, a suíte Pentaho [4].

Apesar de ser possível popular data marts manualmente, escrevendo scripts que leiam de planilhas e insiram no banco de dados, existem diversas ferramentas feitas para facilitar essa etapa, que é a mais trabalhosa na criação de um data warehouse. Escolhemos a ferramenta Spoon [4], do subconjunto Kettle da suíte Pentaho.

Existem dois tipos de scripts que podem ser desenvolvidos no Kettle: transformações e trabalhos. As próximas seções esclarecem a diferença entre eles, com exemplos extraídos do projeto.

3.4.1 Transformações

Transformações manipulam os dados em baixo nível, linha por linha, no mais amplo sentido de extração, transformação e carregamento [3]. Elas são formadas por passos, que são tarefas básicas como ler de um arquivo, filtrar linhas, limpar dados, carregar os dados em um banco, entre muitas outras.

Como passos são tão simples por natureza, mesmo uma transformação pequena pode ser composta de diversos passos. Por exemplo, vejamos a transformação que preenche a Dimensão Professor, Figura 7. O primeiro passo é adicionar um valor padrão que será usado sempre que não soubermos qual professor estava associado ao fato preenchido. A lista de professores está disponível em uma planilha chamada "Histórico das Disciplinas", portanto o segundo passo é ler todas as linhas dessa planilha. A única informação que estamos interessados em é o nome do professor, portanto selecionamos essa única coluna no passo 3. Como o mesmo professor pode aparecer diversas vezes nesta planilha, ordenamos as linhas alfabeticamente e removemos todas as duplicatas. Em seguida, precisamos nos certificar de que a formatação do texto está de acordo com nossos padrões, nos passos 5 e 6: nenhum

espaço em branco antes ou depois do nome, nomes devem estar em caixa baixa e capitalizados. Finalmente, com a lista extraída e transformada, é hora de carregá-la no banco de dados, no passo 7. Com isso, a tabela de professores está devidamente preenchida no banco.

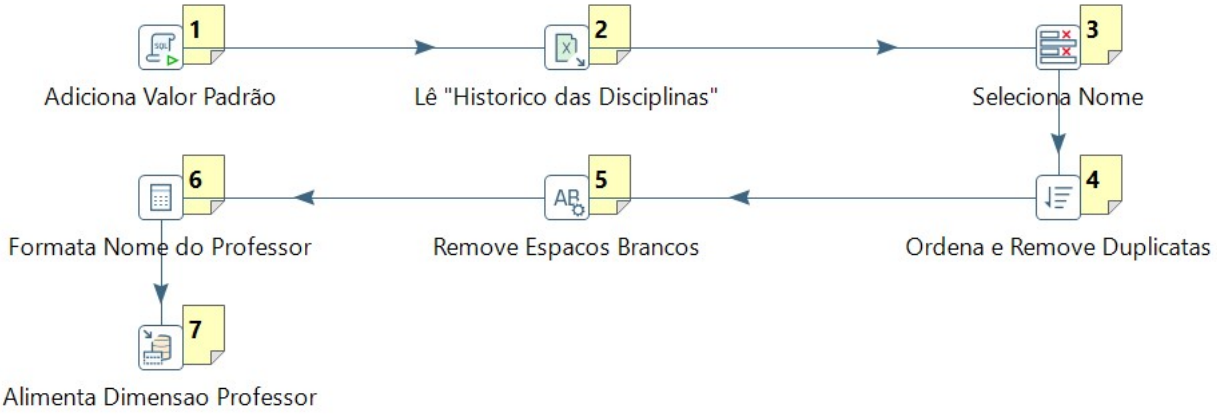


Figura 7: Transformação da Dimensão Professor.

Todas as dimensões podem ser preenchidas por transformações como esta. Algumas exigem a combinação de diversas planilhas (Dimensão Aluno) enquanto outras podem ser geradas sem nenhuma entrada (Dimensão Dia da Semana). Fatos, porém, são inerentemente complexos. Para facilitar o processo de alimento das tabelas fato, podemos dividí-lo em etapas. Isso não é possível só com transformações, pois passos são paralelos. Para reforçar a ordem de execução, vamos precisar de trabalhos.

3.4.2 Trabalhos

Trabalhos são um conjunto ordenado de transformações e/ou trabalhos. Eles não são paralelos, garantindo uma sequência de execução.

Um exemplo simples de trabalho pode ser visto na Figura 8. Nele, estamos interessados em preencher a tabela ponte de professores. Essa tabela possui duas chaves estrangeiras: a chave do professor e a chave do grupo de professores. A dimensão professor já foi preenchida, mas a dimensão grupo de professores ainda não. Por conta da restrição de chave estrangeira, a dimensão grupo precisa ser completamente preenchida antes da tabela ponte. Portanto, podemos criar um

trabalho com apenas dois passos: no primeiro, preenchemos a dimensão grupo; no segundo, atualizamos a tabela ponte.

Figura 8: Trabalho da Ponte Professores.

Pontes e fatos são mais facilmente preenchíveis quando organizados em trabalhos. Uma vez que cada tabela da banco tenha uma transformação/trabalho correspondente que a popula, é possível juntar todo o preenchimento do banco em um único trabalho de fácil entendimento, representado pela Figura 9.

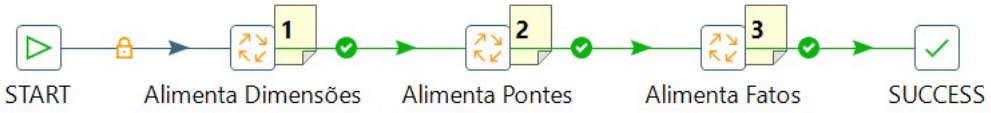


Figura 9: Trabalho que popula o banco por completo.

Com o banco modelado, construído e populado, nosso ambiente DW está preparado para análise. O próximo capítulo aborda consultas definidas e resultados encontrados.

4 ANÁLISES

Neste capítulo vamos observar as análises feitas a partir de nossos data marts e teorizar possíveis implicações.

4.1 CONSIDERAÇÕES GERAIS

Os dados analisados começam nos anos 2000 e vão até 2017. Cabe ressaltar que, em 2010, um novo currículo foi adotado pelo curso de Ciência da Computação. Para fins deste estudo, todo aluno matriculado a partir de 2009 terá seu desempenho avaliado baseado no currículo novo, visto que a maioria dos ingressantes de 2009 entraram com processo para troca de currículo.

A forma de ingresso dos alunos também sofreu alterações durante o período estudado. O vestibular da UFRJ foi aplicado pela última vez em 2011, contabilizando 40% das vagas, sendo outros 40% destinados ao SiSU e 20% a cotas. No segundo semestre, o vestibular foi suspenso e a única forma de ingresso passou a ser o SiSU.

4.2 DESEMPENHO NO PRIMEIRO PERÍODO

Para começar, vamos observar o período de ingresso de cada aluno. O primeiro período na grade recomendada é quase o mesmo para ambos os currículos. Cálculo Infinitesimal I, Computação I, Fundamentos da Computação Digital, Números Inteiros e Criptografia são as disciplinas compartilhadas, sendo que esta última ganhou 30 horas práticas no novo currículo. O currículo antigo incluía Cálculo Vetorial e Geometria Analítica, que foi substituído por Sistemas de Informação.

A granularidade de nosso estudo será o desempenho médio, por período de ingresso, em cada disciplina associada ao primeiro período. Para esse estudo, separamos em dois grupos: calouros e veteranos. Calouros são aqueles que acabaram de ingressar na faculdade e estão fazendo a disciplina pela primeira vez. Veteranos entraram antes daquele período e reprovaram ou trancaram a disciplina anterior-

mente.

4.2.1 Conceito Médio

As figuras 10 e 11 mostram a situação em Computação I para os currículos antigo e novo, respectivamente. De imediato percebemos que o desempenho de calouros é melhor do que o de veteranos em quase todos os períodos. Isso é verdade para todas as disciplinas do primeiro período. Também é possível perceber um padrão interessante: em geral, os alunos do segundo semestre tem pior desempenho. Além disso, as médias vinham caindo durante a vigência do currículo antigo, e não tiveram melhora significativa com o currículo novo.

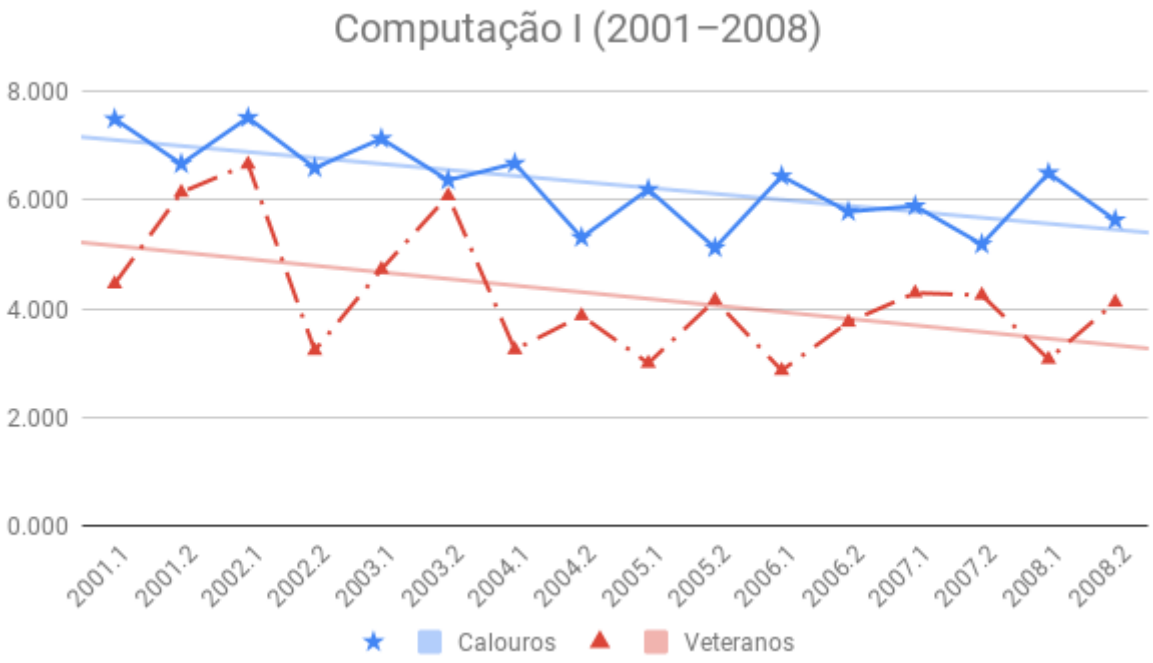


Figura 10: Currículo Antigo - Computação I (2001–2008)

De fato, este padrão decrescente para o currículo antigo e crescente para o currículo novo também pode ser observado em outras disciplinas, como Cálculo Infinitesimal I e Fundamentos da Computação Digital. Gráficos relacionados podem ser encontrados no Apêndice 5.3.

No novo currículo, Números Inteiros e Criptografia teve adição de 30 horas prá-

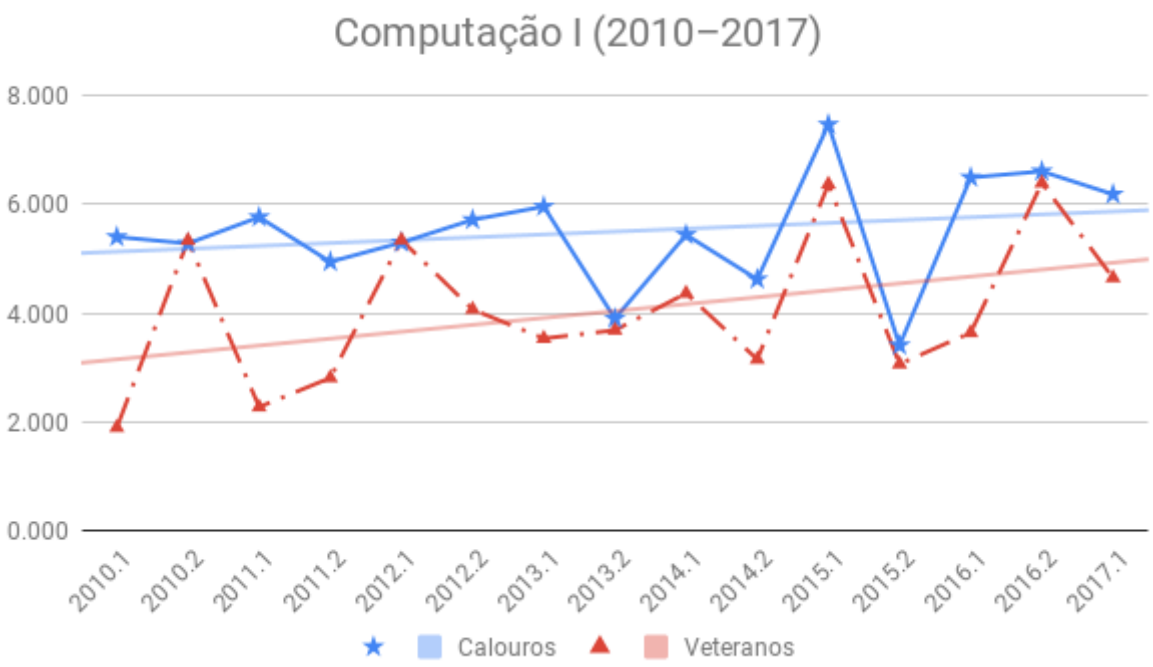


Figura 11: Currículo Novo - Computação I (2010–2017)

ticas. O desempenho dos alunos nesta disciplina vem decaindo significamente desde 2002, como podemos ver nas figuras 12 e 13. Calouros demonstram uma tendência a se estabilizar na média necessária para aprovação, mas veteranos estão em declínio.

4.2.2 Situação Final

Outro ponto de vista interessante é a proporção de aprovações, reprovações e trancamentos para as disciplinas do primeiro período. Assim como na análise de conceito médio, observamos cada disciplina separando os alunos em calouros e veteranos.

Na Figura 14 temos um gráfico de pizza para a situação dos alunos em Números Inteiros e Criptografia pelo currículo antigo. Podemos notar de imediato que o trancamento é mais frequente entre veteranos, quase inexistente para calouros. Calouros também tem menos reprovações por média e falta, o que confirma a percepção geral de que veteranos são mais propensos a faltar aulas. Finalmente, a proporção de calouros aprovada é ligeiramente maior.

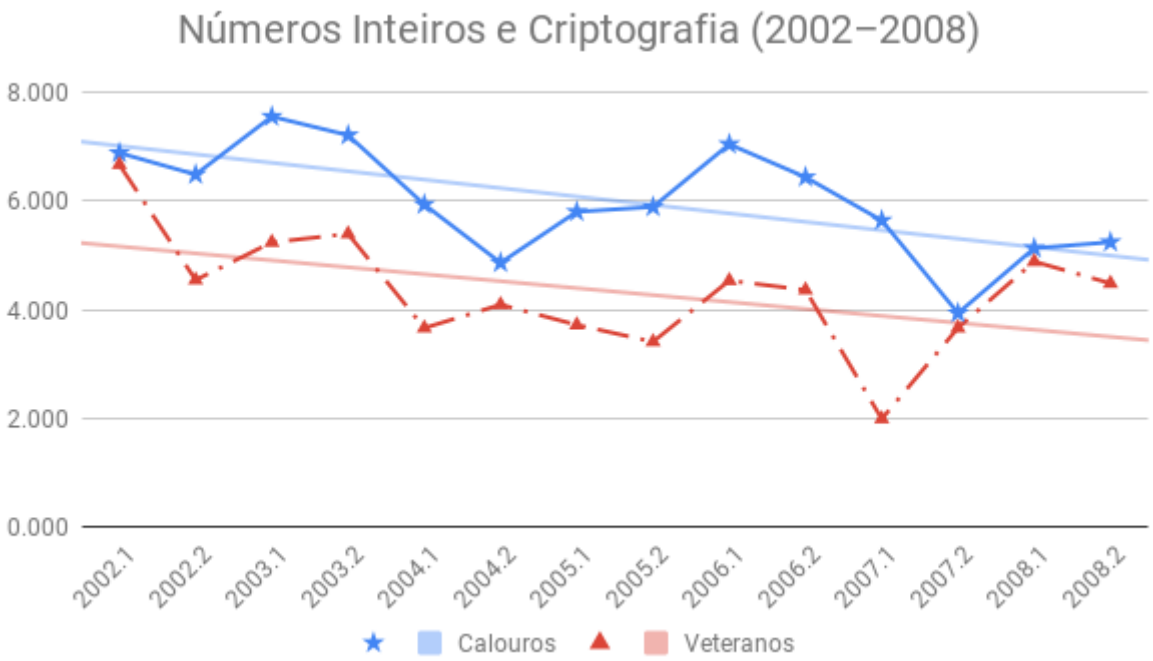


Figura 12: Currículo Antigo - Números Inteiros e Criptografia (2002–2008)

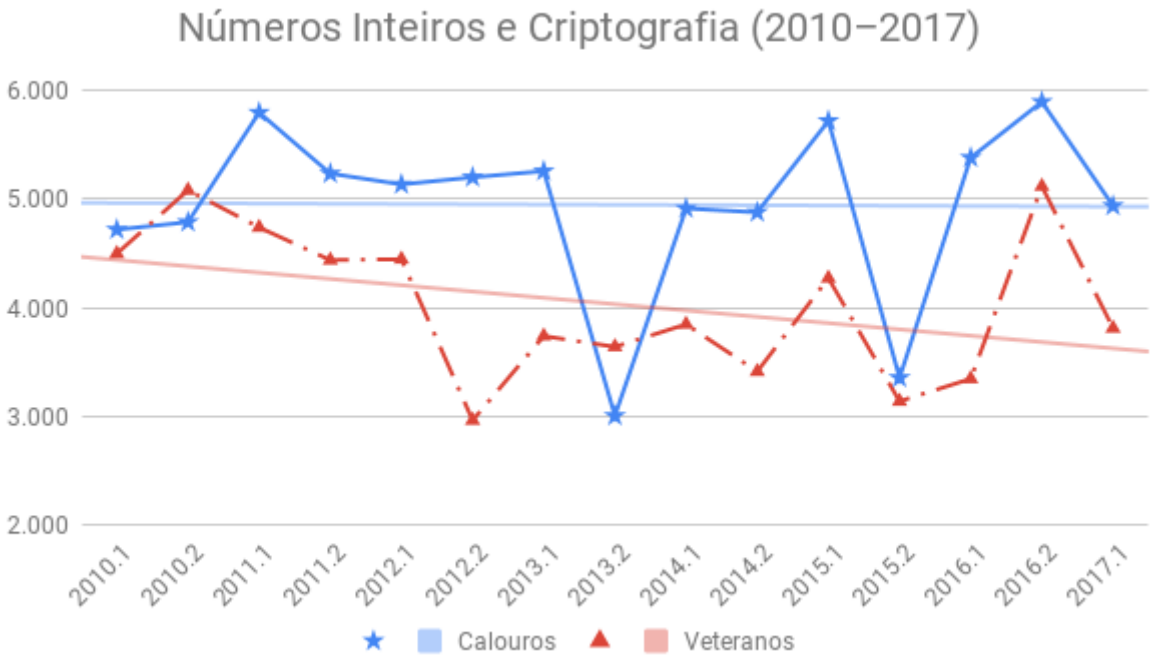


Figura 13: Currículo Novo - Números Inteiros e Criptografia (2010–2017)

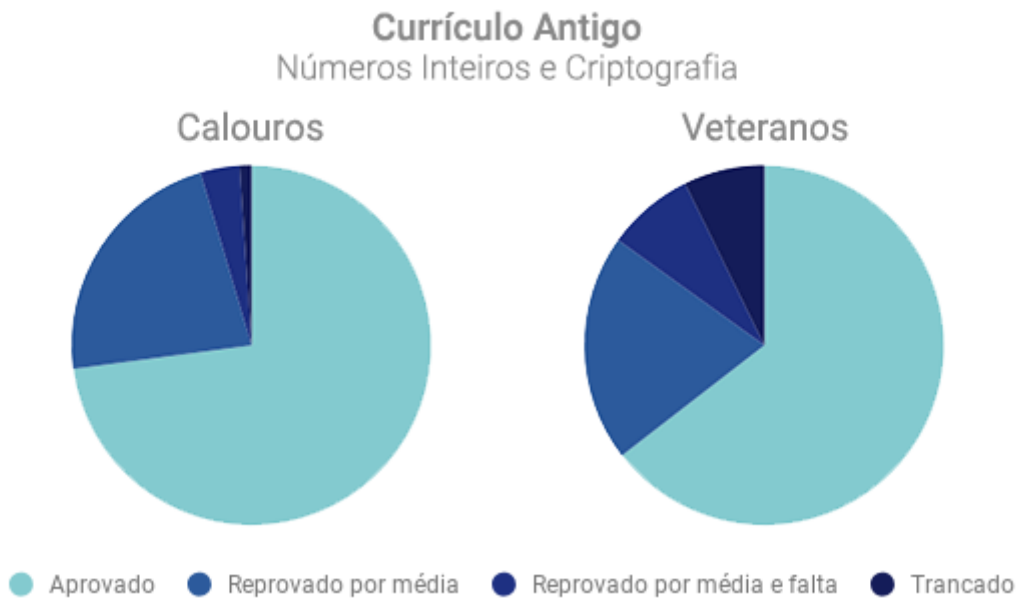


Figura 14: Currículo Antigo - Situação em Números Inteiros e Criptografia

De fato, o padrão descrito é verdadeiro para todas as disciplinas do primeiro período, independente do currículo. A Figura 15 mostra o gráfico correspondente para o currículo novo. Note que a comparação entre calouros e veteranos se mantém, porém existe uma disparidade entre os currículos quando se trata da proporção de aprovações: tanto calouros quanto veteranos tiveram desempenho pior no currículo novo. Este mesmo resultado pode ser observado para todas as disciplinas analisadas, cujos gráfico estão disponíveis no Apêndice 5.3.

Essa discrepância não está relacionada aos currículos em si. É importante ressaltar que pouco depois da troca de currículo, aconteceu uma grande mudança na forma de ingresso dos alunos, com a adoção do SiSU. A queda no desempenho pode indicar despreparo dos ingressantes mais recentes. Além disso, o processo de mudança de curso foi facilitado, possivelmente removendo uma parte da pressão psicológica sentida por alunos que decidiam trocar de curso. Sendo assim, estes alunos hesitariam menos em trancar ou mesmo abandonar disciplinas

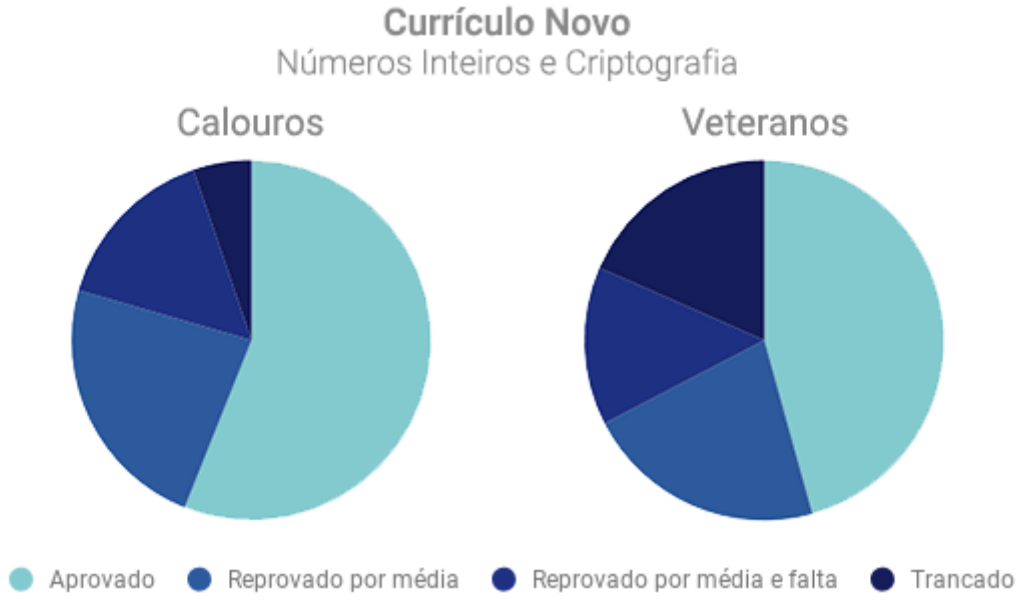


Figura 15: Currículo Novo - Situação em Números Inteiros e Criptografia

4.3 CONCLUSÃO DOS PRIMEIROS ANOS

Os dois primeiros anos de curso são essenciais para a formação do aluno. As disciplinas recomendadas neste tempo constituem uma espécie de ciclo básico da Ciência da Computação. Contudo, é muito comum que os alunos evitem a grade recomendada e escolham fazer disciplinas fundamentais, como Computadores e Programação, apenas no final do curso.

Nesta análise, vamos observar quanto tempo em média os alunos levam para concluir todas as disciplinas referentes aos quatro primeiros períodos da graduação. [1] Também observamos, para aqueles que não foram aprovados, quanto tempo perseveraram em tentativas.

Para cada aluno que recebeu aprovação em todas as disciplinas dos quatro primeiros períodos, observamos qual foi o último período no qual ele esteve inscrito nas mesmas, e calculamos a diferença entre sua entrada na graduação e este período encontrado. O ideal, seguindo a grade recomendada e sem reprovações, é de uma

[1] Para o currículo antigo, desconsideramos a disciplina Computação II, pois a maioria dos alunos não tinha nenhum registro de inscrição na mesma, indicando incompletude dos dados recebidos.

diferença de 4 períodos.

A Tabela 3 mostra as estatísticas encontradas para os currículos antigo e novo. Apesar do número de casos observados ser maior para o currículo antigo, as médias são muito próximas. Pelo desvio padrão, sabemos que cerca de 68% dos casos observados estarão entre 5 e 12 períodos para o currículo antigo, e 5 e 10 períodos para o currículo novo. Em outras palavras, a maioria dos alunos leva entre dois e quatro anos para completar os quatro primeiros períodos.

Estatística	**Currículo Antigo**	**Currículo Novo**
Média	8,18	7,21
Mediana	8	7
Moda	4	7
Desvio Padrão	3,34	2,45
Proporção no Prazo Esperado	13,10%	11,24%
Proporção sem Reprovações	16,59%	15,98%

Tabela 3: Estatísticas - Tempo de Compleção dos Quatro Primeiros Períodos

Os histogramas das figuras 16 e 17 mostram a similaridade da distribuição dos dois currículos, ambos com cauda à direita. O currículo antigo possui valores atípicos maiores, mas é possível que esta diferença se deva ao fato de que os dados só vão até 2017.

No currículo novo, a compleção dos quatro primeiros períodos é obrigatória para que um aluno possa ter seu estágio atestado pela universidade. A pequena tendência a maior atraso no currículo novo comparado ao antigo pode indicar que esta mudança ainda não se refletiu na decisão dos alunos ao montar suas grades curriculares.

Quanto aos alunos que não completaram os quatro primeiros períodos, a Tabela 4 mostra estatísticas análogas àquelas encontradas para alunos concluintes. As médias entre o currículo antigo e o novo também são muito parecidas para esses alunos, com

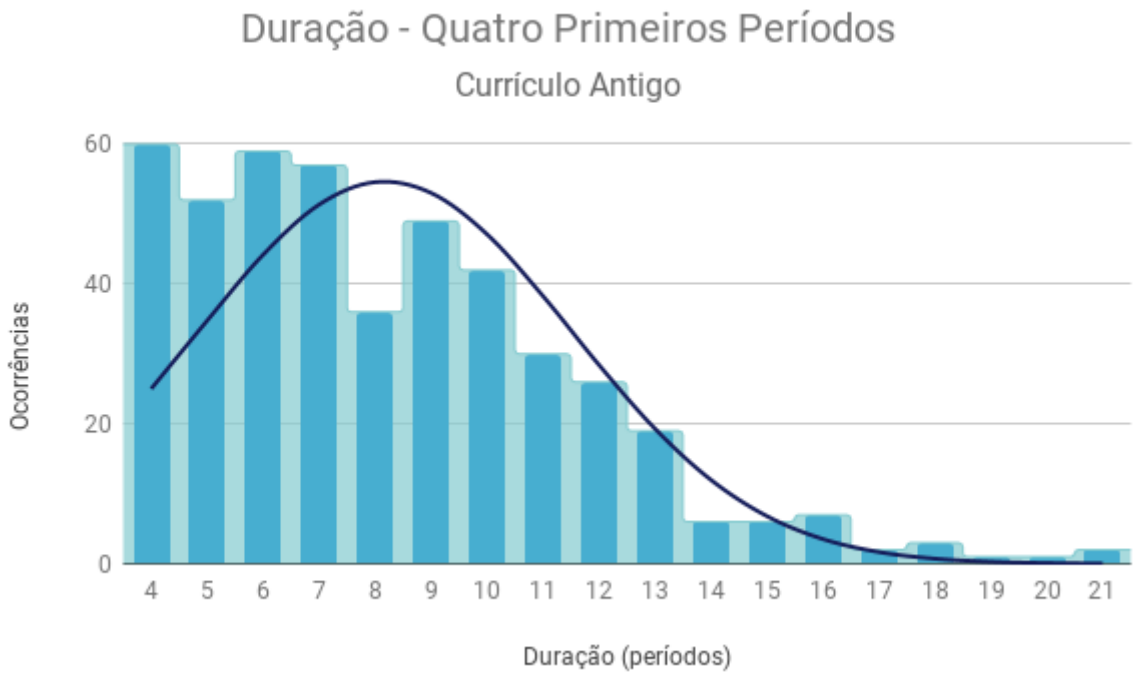


Figura 16: Currículo Antigo - Compleção dos Quatro Primeiros Períodos

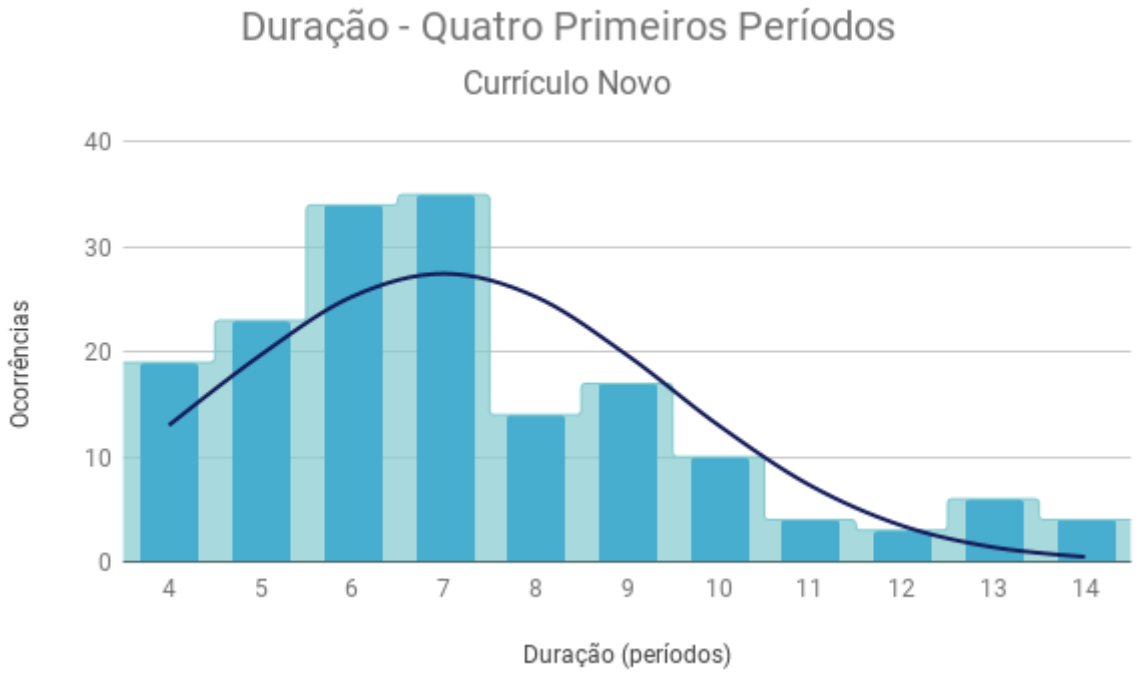


Figura 17: Currículo Novo - Compleção dos Quatro Primeiros Períodos

68% dos casos estando entre 9 e 14 períodos para o antigo e 10 e 13 para o currículo novo. Assim, a maioria dos alunos continua tentando concluir os quatro primeiros períodos por quatro a seis anos.

Estatística	**Currículo Antigo**	**Currículo Novo**
Média	7	5,97
Mediana	7	6
Moda	1	1
Desvio Padrão	4,47	3,61
Proporção sem Terminar	54,56%	74,55%
Proporção sem Reprovações	3,27%	4,24%
Conclusão Média	53,99%	47,14%

Tabela 4: Estatísticas - Tempo Tentando os Quatro Primeiros Períodos

É interessante notar que alguns dos alunos não tem reprovações; estes nem sequer chegaram a se inscrever em todas as disciplinas. Alguns ainda tem matrícula ativa, mas muitos deixaram o curso no primeiro período - o que explica a moda ser 1 para ambos os currículos. A proporção de alunos que não terminaram os quatro primeiros períodos é maior para o currículo novo, contudo isso pode ser devido aos alunos mais recentes ainda não terem tido oportunidade, afinal os resultados anteriores mostram que é comum levar cerca de 8 períodos para obter aprovação nestas disciplinas.

4.4 EVASÃO

Com a introdução do SiSU como forma de ingresso, a flexibilidade na mudança de curso e a mudança de currículo, cabe o questionamento de se o número de evasões tem crescido nos últimos anos. As análises a seguir procuram retratar o cenário de evasões do curso de Ciência da Computação.

Houve uma limitação técnica com relação aos resultados. Como os dados recebi-

dos se diziam somente a alunos matriculados em Ciência da Computação, dados de alunos que porventura fizeram transferência interna não estão presentes em nosso banco. Evidentemente, a incompletude dos dados prejudica a corretude das análises. Melhores resultados poderiam ser obtidos caso todos os alunos que um dia fizeram parte do curso fossem considerados.

Por esse motivo, os números apresentados a seguir são estimativas *otimistas* da evasão do curso.

4.4.1 Quatro Primeiros Períodos

Espera-se que o maior número de desistências aconteça dentro dos dois primeiros anos do aluno. Para observar a evasão no início do curso, calculamos quantos alunos estiveram inscritos em apenas um, dois, três ou quatro períodos. Os resultados foram dispostos em quatro gráficos diferentes, apresentados na Figura 18. Na Figura 19 os mesmos dados são representados em forma de diagramas de caixa, para fácil comparação entre períodos e identificação de valores atípicos.

Como esperado, o primeiro período concentra o maior número de evasões, com uma média de 12 desistências por ano. O segundo e o terceiro período aparecem empatados com cerca de 6 desistências por ano, e o quarto período tem o menor resultado: 4 desistências por ano.

Desconsiderando valores atípicos, não foi possível observar nenhuma tendência de maior evasão nos últimos anos.

4.4.2 Retenção de Alunos

Uma boa forma de visualizar a retenção de alunos no curso é através de um gráfico de funil, visto na Figura 20. Para cada período, observamos os alunos que ingressaram no mesmo ao longo dos anos, verificando quantos destes ainda estão ativos. Calculamos, então, a média de alunos ativos a cada período para ter uma visão geral da evasão do curso. Assim como nas análises anteriores, fica claro que o

Figura 18: Evasão nos Quatro Primeiros Períodos - Colunas

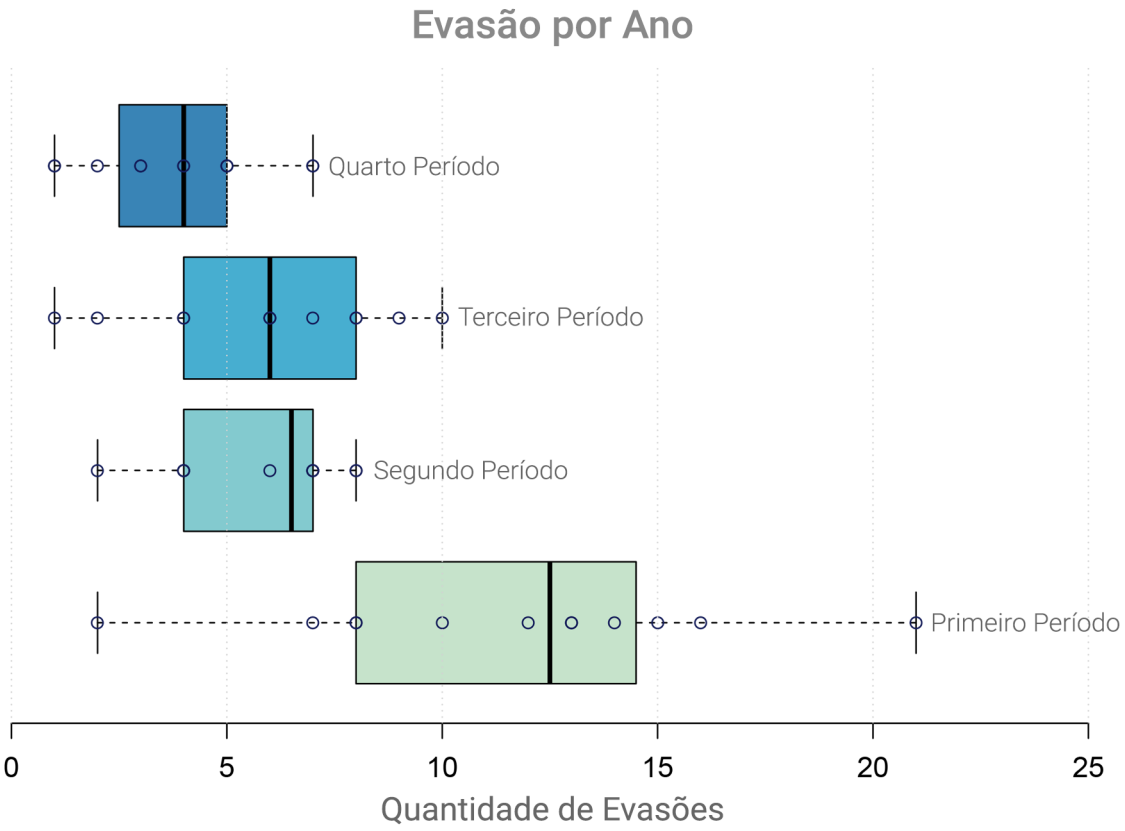


Figura 19: Evasão nos Quatro Primeiros Períodos - Caixas

período onde mais perdemos alunos é no primeiro período.

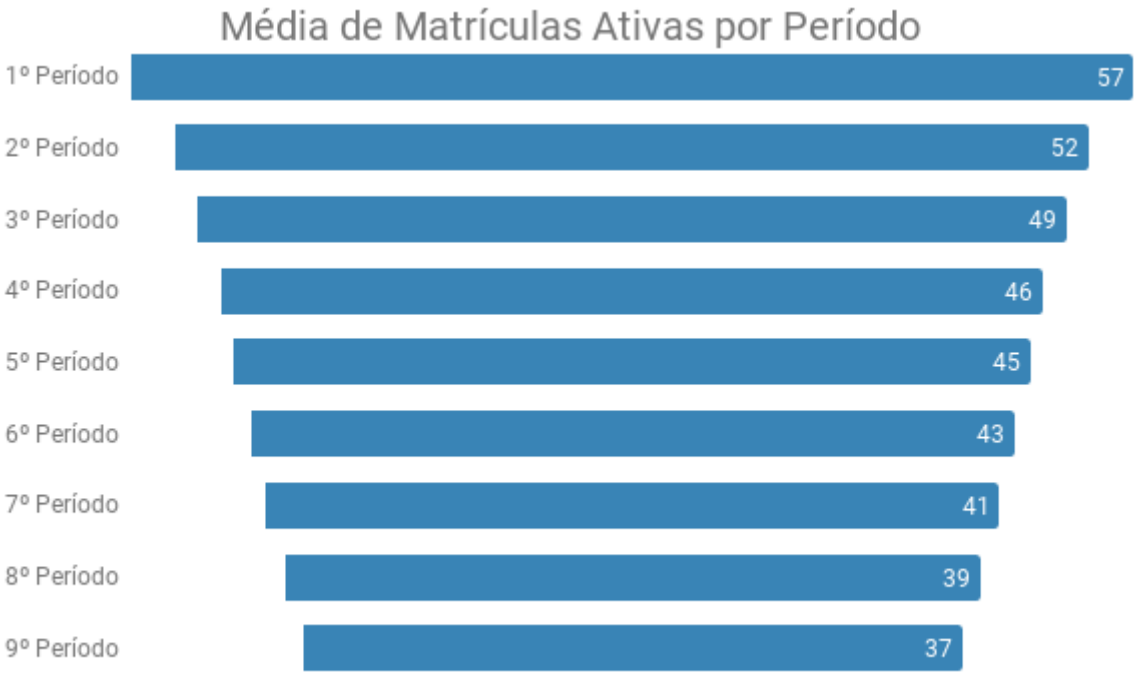


Figura 20: Média de Matrículas Ativas por Período

4.5 CONCLUSÃO E ABANDONO

Nossa última análise observa qual a proporção de alunos que chega ao final do curso. Queremos saber quanto tempo, em média, é necessário para graduar.

Atividades Complementares e Projeto Final não estavam entre as disciplinas cursadas nos dados recebidos. Por esse motivo, consideramos concluintes os alunos que tenham completado todas as disciplinas obrigatórias de seu currículo. Para um aluno ser marcado como abandono, ele não é concluinte e não esteve inscrito em nenhuma disciplina nos três últimos períodos avaliados (2016.1, 2016.2 e 2017.1).

A Tabela 5 mostra as estatísticas do tempo levado para concluir o curso, calculado como a diferença entre o período de ingresso e o período de conclusão. Em média, tanto para o currículo novo quanto para o antigo, os alunos costumam levar 11 períodos para completar o curso, um ano a mais na faculdade do que o esperado.

Estatística	Currículo Antigo	Currículo Novo
Média	11,63	11,37
Mediana	11	11
Moda	9	11
Desvio Padrão	2,90	1,68
Mais Curto	7	8
Mais Longo	21	16

Tabela 5: Estatísticas - Tempo para Concluir o Curso (períodos)

A Figura 21 mostra a quantidade de alunos que ingressaram, concluíram e abandonaram o curso para cada ano. É possível perceber uma subida considerável de abandonos em 2008 e 2009. Os dados para estes anos tem uma margem de erro maior, uma vez que os alunos ingressantes dos mesmos cursaram uma mistura de currículos, dificultando a identificação de conclusão do curso.

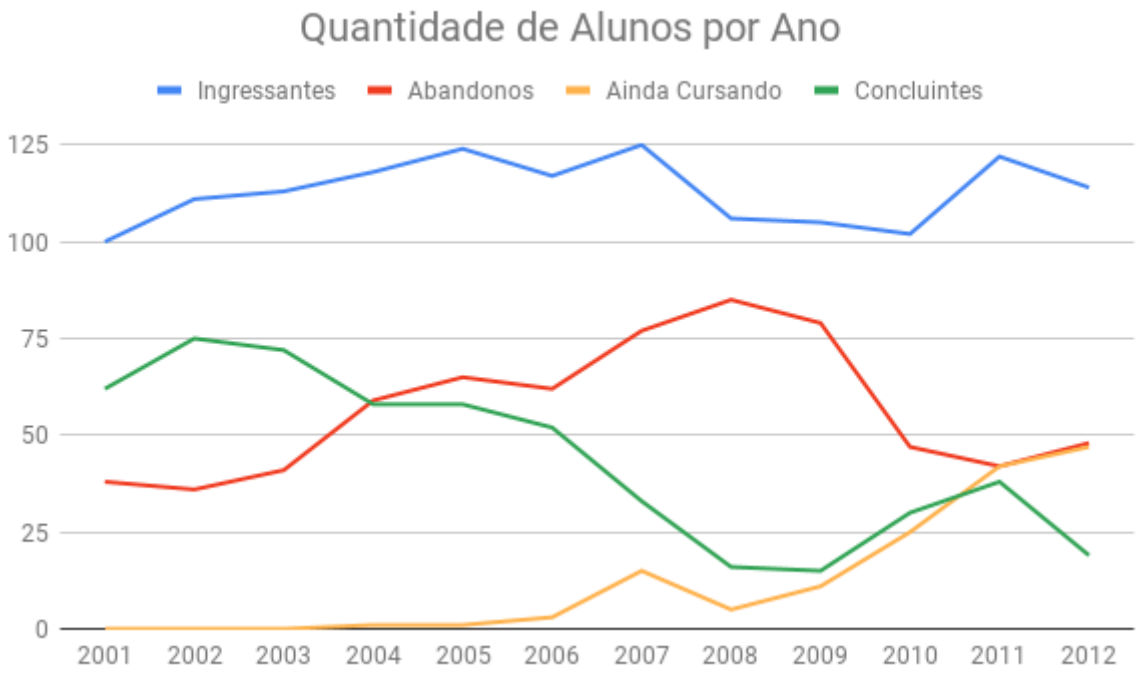


Figura 21: Quantidade de Alunos por Ano

Outra forma de ver os mesmos dados é através de um gráfico de áreas, represen-

tado na Figura 22. Cada área representa a proporção dos alunos que concluíram/abandonaram/ainda estão cursando para cada ano. Podemos perceber que para anos mais distantes, como 2002, não temos mais alunos cursando, e a porcentagem de concluintes é maior. A faixa de abandono é mais estreita nos anos mais recentes, pois parte significativa dos alunos ainda está ativa.

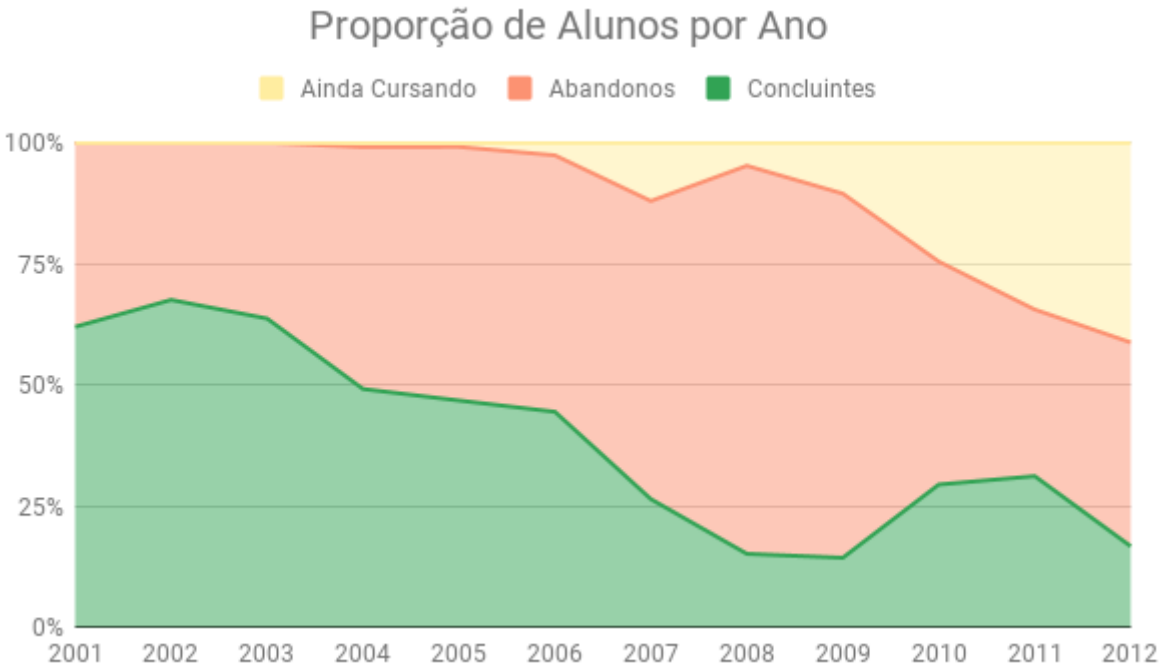


Figura 22: Proporção de Alunos por Ano

5 CONCLUSÃO

Neste trabalho, propomos um modelo dimensional para retratar a situação geral do curso de Ciência da Computação na UFRJ. Este modelo foi implementado, produzindo um banco de dados analítico. O banco foi populado utilizando dados enviados pelo SIGA, além de dados extraídos manualmente da plataforma. Uma vez construído o banco, fizemos análises sobre a coleção de dados, mostrando o potencial que um data warehouse tem a oferecer.

5.1 RESULTADOS

Nossos resultados, ainda que primitivos, trazem reflexão sobre o cenário atual do curso. Mostramos que a queda no desempenho dos alunos data de antes da mudança de currículo e da adoção do SiSU; apesar disso, a quantidade de reprovações tem se agravado nos últimos anos. Também observamos que a maioria dos alunos que abandona o curso o faz no primeiro período, e que as evasões decrescem quase linearmente nos períodos seguintes. Finalmente, confirmamos a suspeita de que os alunos levam em média um ano a mais para concluir o curso.

O principal resultado, porém, é a demonstração de que um banco analítico como este pode ser de grande auxílio a professores, dando suporte à tomada de decisão da coordenação. O modelo desenvolvido para esse trabalho pode ser reutilizado no futuro para responder mais perguntas interessantes, como:

- Qual a relação entre a evasão do aluno e seu desempenho até aquele momento?
- Quantas vezes em média os alunos precisam fazer as disciplinas do primeiro período para obter aprovação nelas?
- Qual disciplina nos primeiros quatro períodos o aluno tem maior dificuldade para obter créditos?
- Se desconsiderarmos disciplinas que não são pré-requisito de nenhuma outra, como Eletromagnetismo e Ondas ou Computadores e Programação, quanto

tempo os alunos levam para completar os quatro primeiros períodos?

5.2 PROBLEMAS ENFRENTADOS

Os dados recebidos, apesar de ricos, eram incompletos. Alunos que porventura trocaram de curso não estavam listados, cancelamento por conclusão de curso não foi recebido, entre outros dados ausentes que enfraqueceram as análises.

Grande parte do esforço do trabalho se resumiu em juntar as peças operacionais para montar o quebra-cabeças do banco analítico. Muitos dados tiveram de ser inferidos com base no que tínhamos, ou extraídos de forma não ótima dos sites do SIGA.

Inconstências na suíte Pentaho também impactaram o andamento do trabalho. Versões supostamente compatíveis das diferentes ferramentas não se comunicavam bem, interrompendo o fluxo de produção.

Pelos motivos acima expostos, a parte mais importante do trabalho, que eram as análises de desempenho dos alunos, acabou sendo mais curta do que se planejava inicialmente.

5.3 TRABALHOS FUTUROS

As análises apresentadas neste trabalho foram apenas uma iniciativa que precisa ser continuada de forma a responder mais perguntas com a estrutura desenvolvida. Para trabalhos futuros, seria importante buscar mais fontes de dados, idealmente vindos diretamente do banco operacional. A modelagem atual é facilmente reaproveitável e expansível, ou seja, novas tabelas fato podem ser incluídas usando dimensões já existentes.

Outra possibilidade interessante seria fazer análises focadas em um aluno específico. Como os dados recebidos foram anonimizados, essa abordagem foge do escopo deste trabalho, mas poderia ser muito útil para seu orientador acadêmico.

REFERÊNCIAS

[1] Best of BI. Data modeling & profiling tool: Sql power architect. `http://www.bestofbi.com/page/architect`. Acesso em: 15-08-2018.

[2] Bouman, R., e van Dongen, J. *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*. Wiley Publishing, Inc., 2009.

[3] Casters, M., Bouman, R., e van Dongen, J. *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. Wiley Publishing, Inc., 2010.

[4] Hitachi Vantara. Pentaho. `https://www.hitachivantara.com/go/pentaho.html`. Acesso em: 15-08-2018.

[5] Hoed, R. M. Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação. Dissertação, Universidade de Brasília, Brasília, 2016.

[6] Kimball, R., e Caserta, J. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, 1 ed. Wiley Publishing, Inc., 2004.

[7] Kimball, R., e Ross, M. *The data warehouse toolkit: the complete guide to dimensional modeling*, 2 ed. Wiley Computer Publishing, 2002.

[8] Kimball Group. Kimball dimensional modeling techniques. Retirado do livro The Data Warehouse Toolkit, 3 ed., 2017.

[9] The PostgreSQL Global Development Group. Postgresql: The world's most advanced open source database. `https://www.postgresql.org/`. Acesso em: 15-08-2018.

ANEXOS

ANEXO A – Dados Recebidos

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	0f089a96932c0e5
Carga Horária	Quantidade de horas acumuladas obtidas pelo aluno.	2850

Tabela 6: Carga Horária Obtida Acumulada

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	004a44fc3728c7e
CR	Coeficiente de rendimento acumulado do aluno até determinado período.	7.61509434
Ano	Ano até o qual esse coeficiente foi acumulado.	2013
Período	Período até o qual esse coeficiente foi acumulado.	2

Tabela 7: CR Acumulado por Período

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	0187947abcdc968
Ano	Ano no qual o coeficiente foi obtido.	2000
Período	Período no qual o coeficiente foi obtido.	1
CRA	Coeficiente de rendimento do aluno naquele período.	6.5

Tabela 8: CR do Período

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	004a44fc3728c7e
Créditos Obtidos Acumulados	Quantidade de créditos acumulados pelo aluno.	76

Tabela 9: Créditos Obtidos Acumulados

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	004a44fc3728c7e
Ano	Ano no qual a disciplina foi cursada.	2012
Período	Período no qual a disciplina foi cursada.	1
Código Disciplina	Código identificador da disciplina.	FIS111
Nome Disciplina	Nome da disciplina.	Fisica Experimental I
Situação	Situação final do aluno ao cursar a disciplina.	Aprovado
Conceito	Conceito do aluno após conclusão (0-100).	085

Tabela 10: Disciplinas Cursadas com Nota e Situação Final

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	5fe9d73af710090
Forma Ingresso	Forma de ingresso do aluno no curso.	SiSU - Sistema de Seleção Unificada
Ano	Ano de ingresso do aluno no curso.	2016
Período	Período de ingresso do aluno no curso.	2

Tabela 11: Forma de Ingresso

CAMPO	DESCRIÇÃO	EXEMPLO
Código da Disciplina	Código da disciplina oferecida.	MAB605
Nome da Turma	Nome da turma aberta naquele período.	Recuperação da Informação
Ementa	Ementa da disciplina oferecida.	Variável de acordo com os tópicos oferecidos.
Nome Professor	Nome do(a) professor(a) responsável pela turma.	ADRIANA SANTAROSA VIVACQUA
Ano de Oferta	Ano no qual a disciplina foi oferecida.	2011
Período de oferta	Período no qual a disciplina foi oferecida.	2
Dia da Semana	Dia da semana no qual aulas são ministradas. Múltiplas linhas são utilizadas se há mais de um dia por semana.	Terça
Hora Início	Horário de início das aulas ministradas nesse dia da semana.	13
Hora Fim	Horário de fim das aulas ministradas nesse dia da semana.	15

Tabela 12: Histórico das Disciplinas

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	7610a42b15f140a
Ano	Ano no qual a disciplina foi trancada.	2004
Período	Período no qual a disciplina foi trancada.	1
Código Disciplina	Código da disciplina trancada.	MAB607
Nome Disciplina	Nome da disciplina trancada.	Empreendimento em Informática

Tabela 13: Matérias Trancadas pelos Alunos

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	01d643e8ff79d8c
CR	Coeficiente de rendimento do aluno.	2.7
Ano	Ano no qual o coeficiente do aluno esteve abaixo de 3.	2017
Período	Período no qual o coeficiente do aluno esteve abaixo de 3.	1

Tabela 14: Períodos com CR Menor que 3

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	ef475a51a0b3438
Ano	Ano no qual a matrícula do aluno foi cancelada por abandono.	2004
Período	Período no qual a matrícula do aluno foi cancelada por abandono.	2

Tabela 15: Periodos de Cancelamento por Abandono

CAMPO	DESCRIÇÃO	EXEMPLO
Aluno	Valor em hash que identifica um aluno.	004a44fc3728c7e
Períodos Trancados	Quantidade de períodos trancados do aluno.	6

Tabela 16: Períodos de Trancamento de Matrícula

ANEXO B – Modelagem Dimensional

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_aluno	INT	Chave substituta da dimensão.	239
cod_anonimo	VARCHAR	Valor em hash que identifica um aluno.	5fe9d73af710090
forma_ingresso	VARCHAR	Forma de ingresso do aluno.	SiSU - Sistema de Seleção Unificada
ano_ingresso	VARCHAR	Ano de ingresso do aluno.	2016
periodo_ingresso	VARCHAR	Período de ingresso do aluno.	2

Tabela 17: Dimensão Aluno

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_dia_semana	INT	Chave substituta da dimensão.	4
dia_extenso	VARCHAR	Dia da semana, extenso.	Quinta-feira
dia_abreviado	VARCHAR	Dia da semana, abreviado.	Quinta

Tabela 18: Dimensão Dia da Semana

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_disciplina	INT	Chave substituta da dimensão.	3257
codigo _disciplina	VARCHAR	Código identificador da disciplina.	MAB605
nome _disciplina	VARCHAR	Nome da disciplina.	Recuperação da Informação
codigo _departamento	VARCHAR	Código do departamento ao qual a disciplina está associada.	MAB
nome _departamento	VARCHAR	Nome do departamento ao qual a disciplina está associada.	Departamento de Ciência da Computação
ementa	VARCHAR	Ementa da disciplina.	Variável de acordo com os tópicos oferecidos.
flag_obrigatoria	BOOLEAN	Verdadeiro se a disciplina é obrigatória no curso de Ciência da Computação.	false

Tabela 19: Dimensão Disciplina

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_matricula	INT	Chave substituta da dimensão.	2
situacao	VARCHAR	Situação de matrícula.	Regular
faixa_cr _acumulado	VARCHAR	Faixa que contém o CR acumulado do aluno.	< 1
faixa_cr _periodo	VARCHAR	Faixa que contém o CR do aluno no período.	$1 <= x < 3$

Tabela 20: Dimensão Matrícula

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_periodo	INT	Chave substituta da dimensão.	50
ano	VARCHAR	Ano que compõe o período.	2012
periodo	VARCHAR	Número que identifica o período.	1

Tabela 21: Dimensão Período

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_professor	INT	Chave substituta da dimensão.	210
nome	VARCHAR	Nome do professor.	Silvana Rossetto

Tabela 22: Dimensão Professor

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
pk_situacao _final	INT	Chave substituta da dimensão.	4
situacao	VARCHAR	Situação final do aluno na disciplina.	Aprovado
faixa_conceito	VARCHAR	Faixa que contém o conceito do aluno na disciplina.	5<= x < 7

Tabela 23: Dimensão Situação Final

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
fk_disciplina	INT	Chave estrangeira referente à disciplina associada.	612
fk_situacao _final	INT	Chave estrangeira referente à situação final associada.	8
fk_aluno	INT	Chave estrangeira referente ao aluno associado.	9
fk_periodo	INT	Chave estrangeira referente ao período associado.	58
fk_professores	INT	Chave estrangeira referente ao conjunto de professores associado.	125
fk_horarios	INT	Chave estrangeira referente ao conjunto de horários associado.	106
conceito	DOUBLE	Conceito obtido nas condições indicadas pelas dimensões.	0
creditos _requisitados	INT	Créditos requisitados referentes à disciplina cursada.	4
creditos _obtidos	INT	Créditos obtidos referentes à disciplina cursada.	0
carga_horaria	INT	Carga horária referente à disciplina cursada.	60

Tabela 24: Fato Situação em Disciplina

CAMPO	TIPO	DESCRIÇÃO	EXEMPLO
fk_matricula	INT	Chave estrangeira referente à matrícula associada.	22
fk_aluno	INT	Chave estrangeira referente ao aluno associado.	373
fk_periodo	INT	Chave estrangeira referente ao período associado.	50
cr_acumulado	DOUBLE	Coeficiente de rendimento acumulado nas condições indicadas pelas dimensões.	5.24
cr_periodo	DOUBLE	Coeficiente de rendimento do período nas condições indicadas pelas dimensões.	6.4

Tabela 25: Fato Situação em Período

ANEXO C – Análises

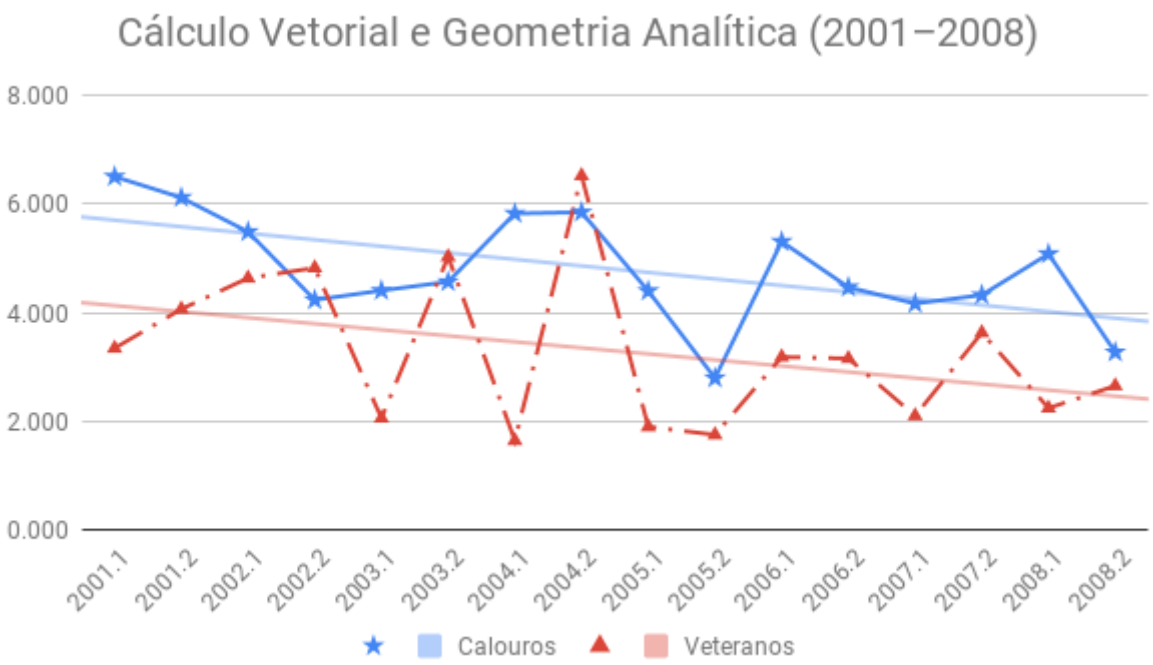


Figura 23: Currículo Antigo - Cálculo Vetorial e Geo. Analítica (2001–2008)

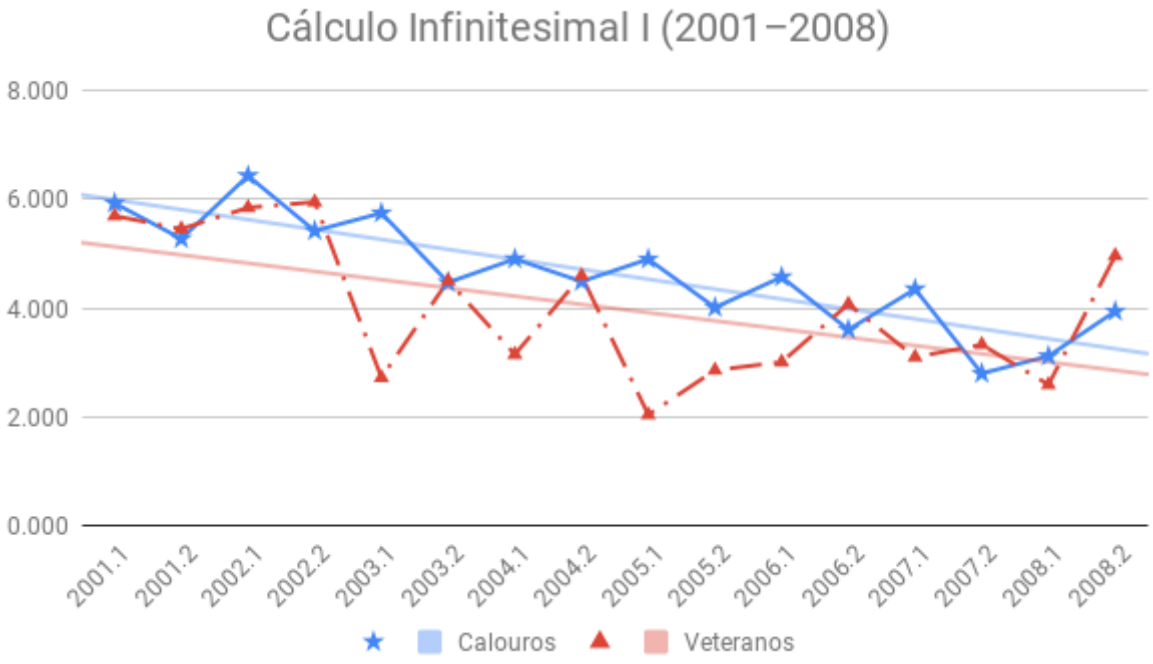


Figura 24: Currículo Antigo - Cálculo Infinitesimal I (2001–2008)

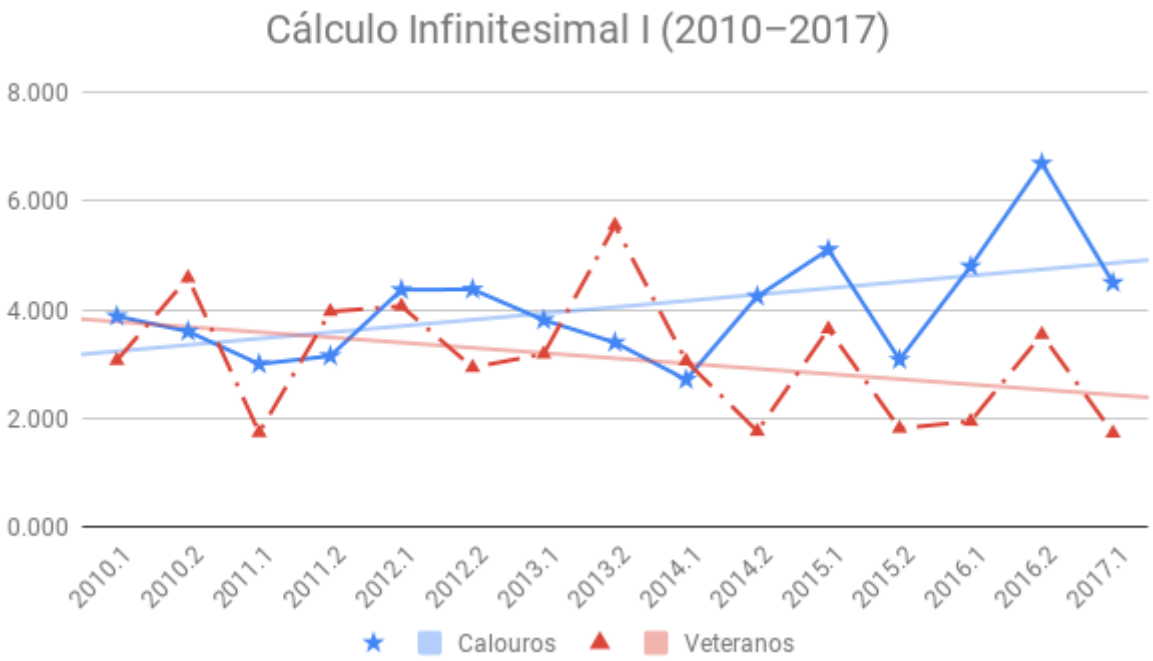


Figura 25: Currículo Novo - Cálculo Infinitesimal I (2010–2017)

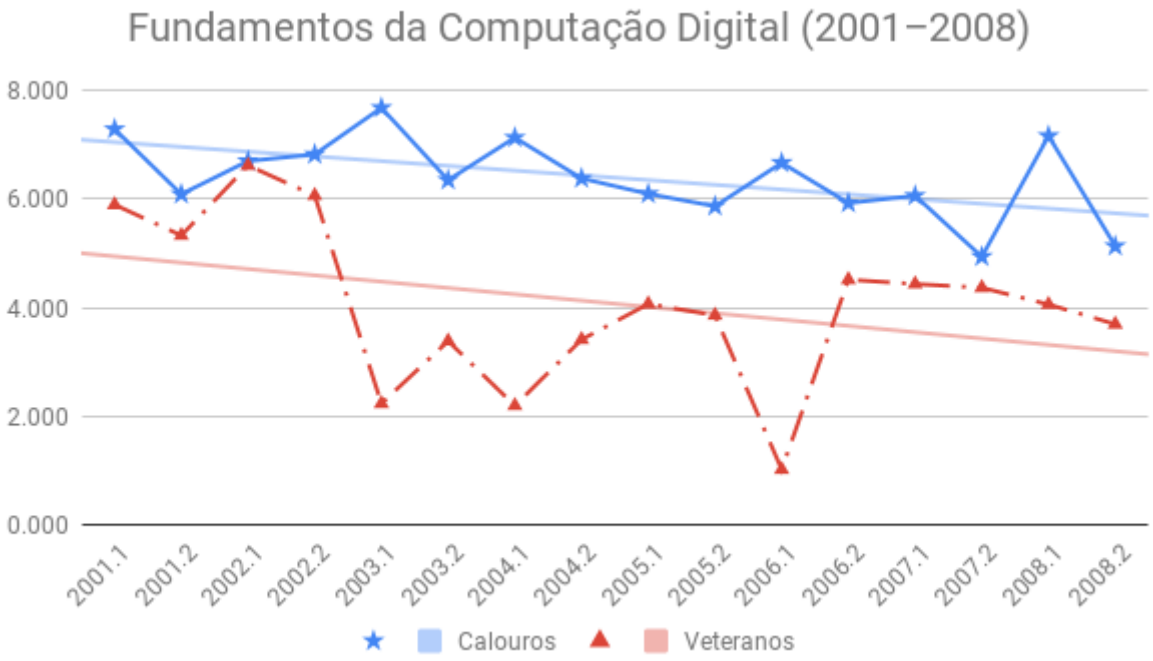


Figura 26: Currículo Antigo - Fundamentos da Computação Digital (2001–2008)

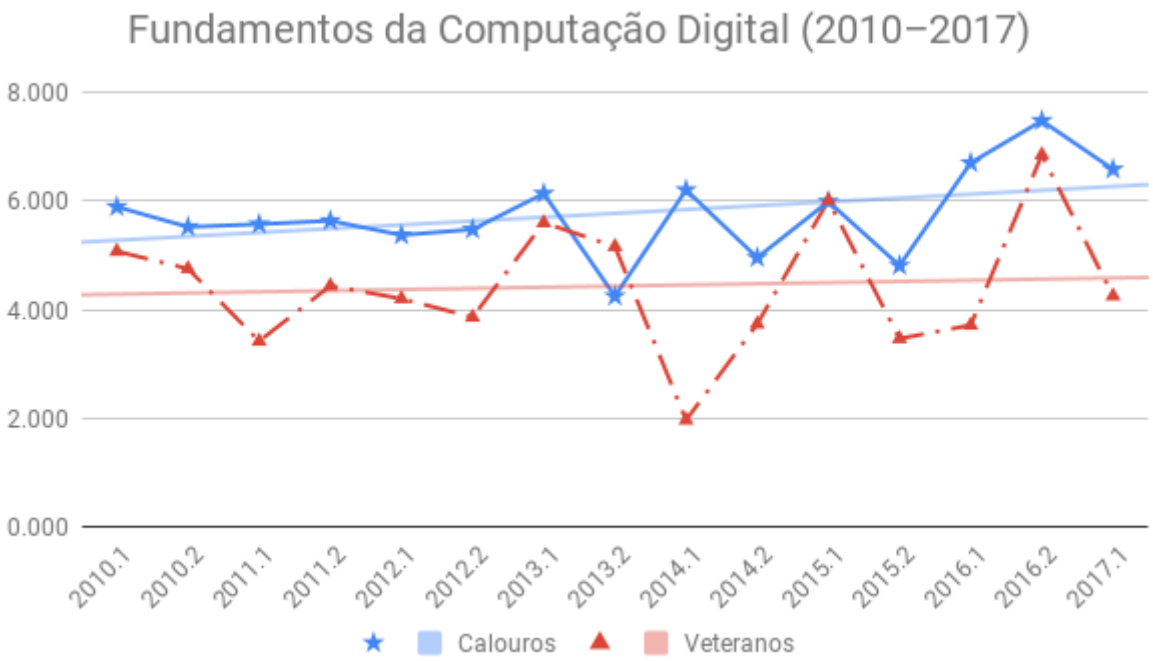


Figura 27: Currículo Novo - Fundamentos da Computação Digital (2010–2017)

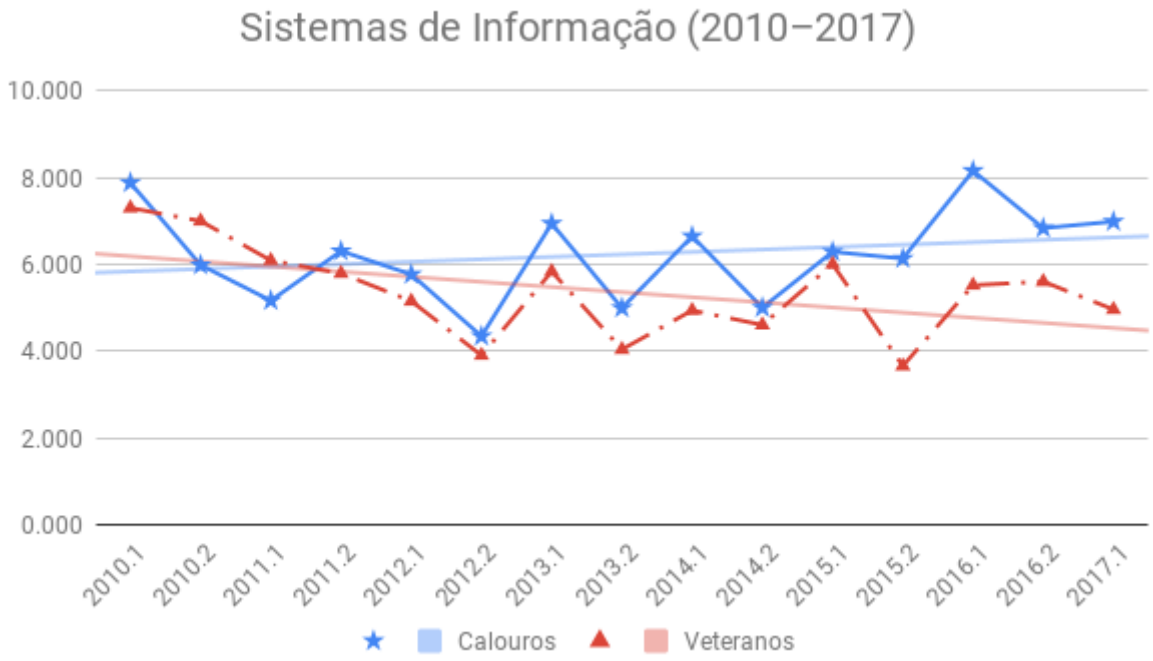


Figura 28: Currículo Novo - Sistemas de Informação (2010–2017)

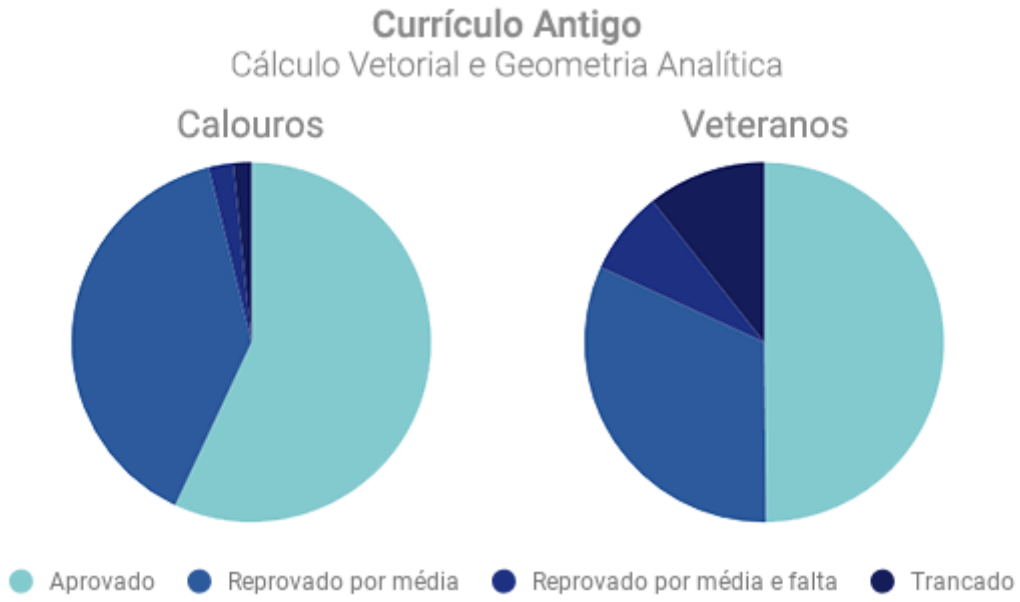


Figura 29: Currículo Antigo - Situação em Cálculo Vetorial e Geo. Analítica

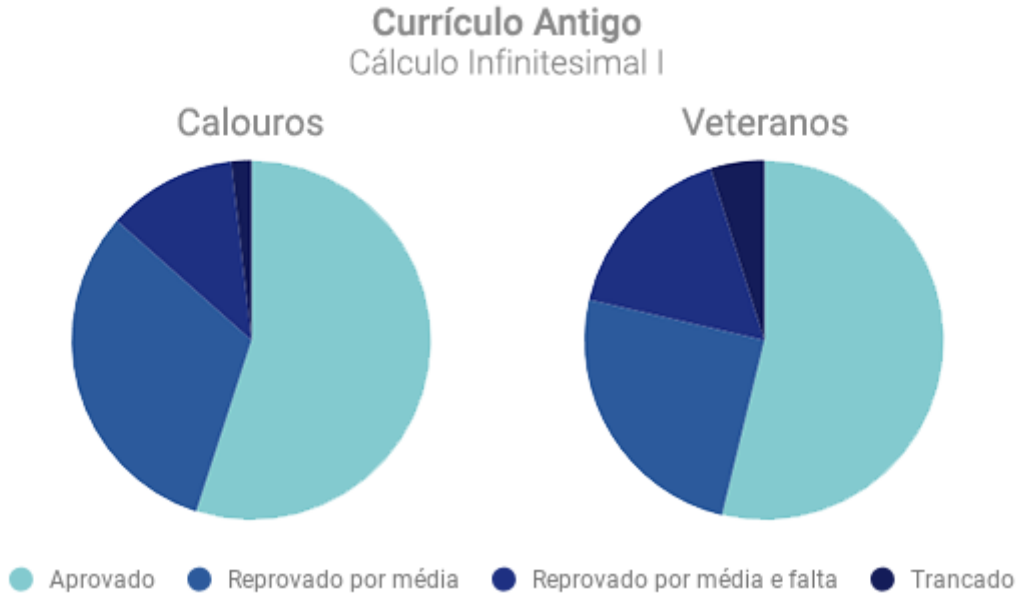


Figura 30: Currículo Antigo - Situação em Cálculo Infinitesimal I

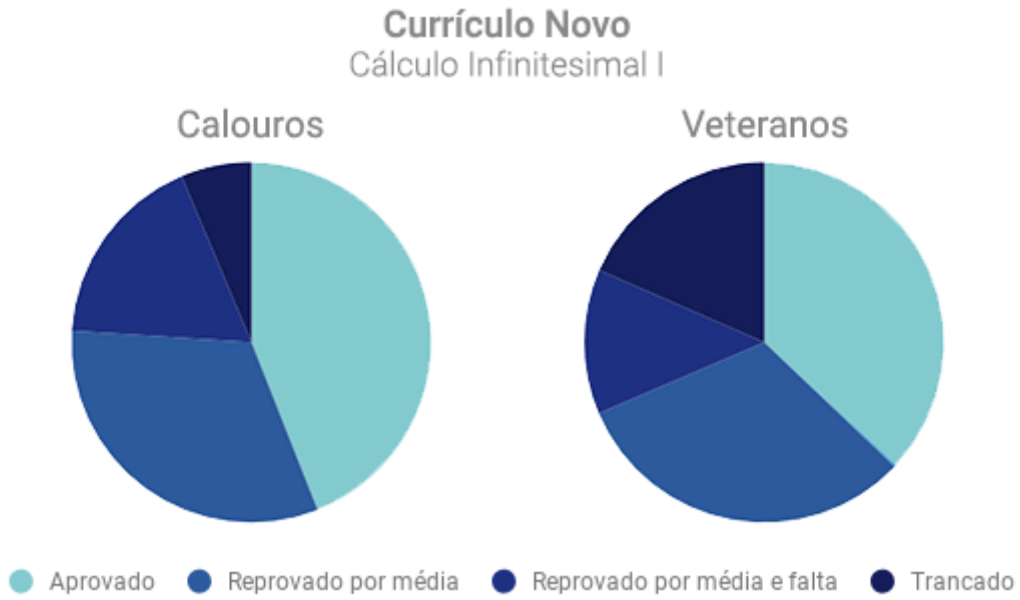


Figura 31: Currículo Novo - Situação em Cálculo Infinitesimal I

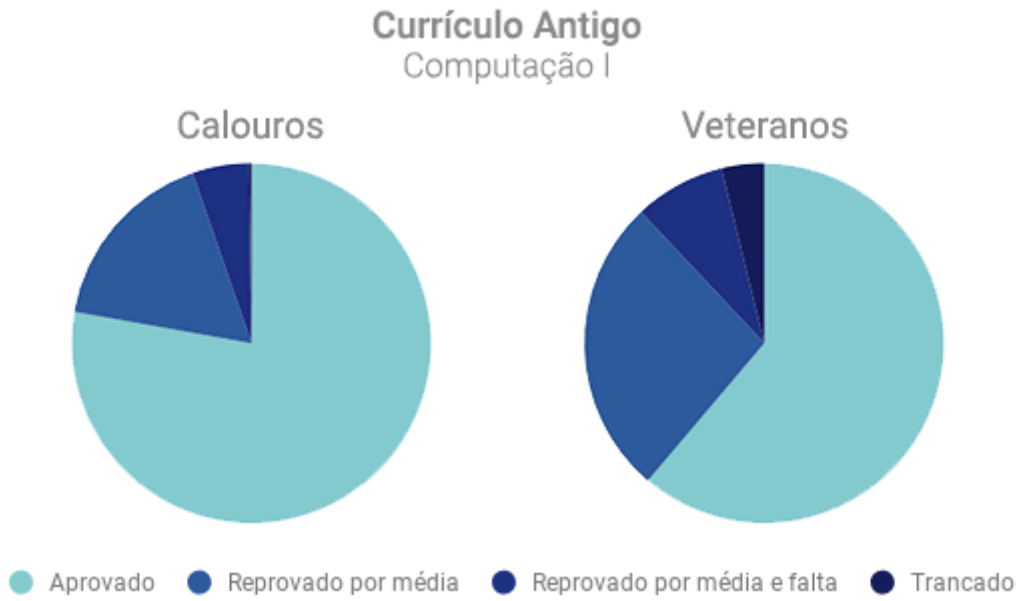


Figura 32: Currículo Antigo - Situação em Computação I

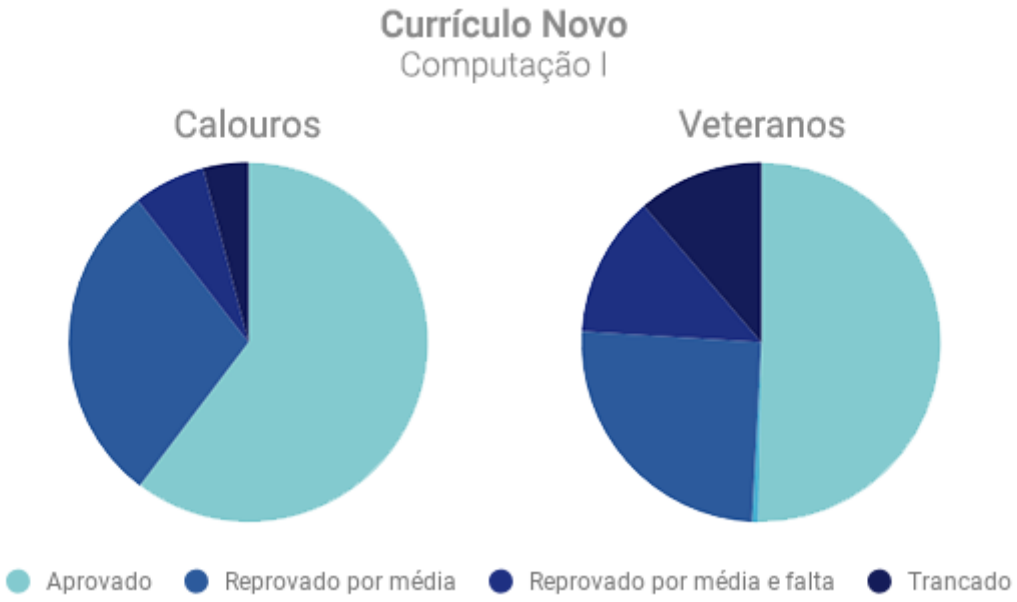


Figura 33: Currículo Novo - Situação em Computação I

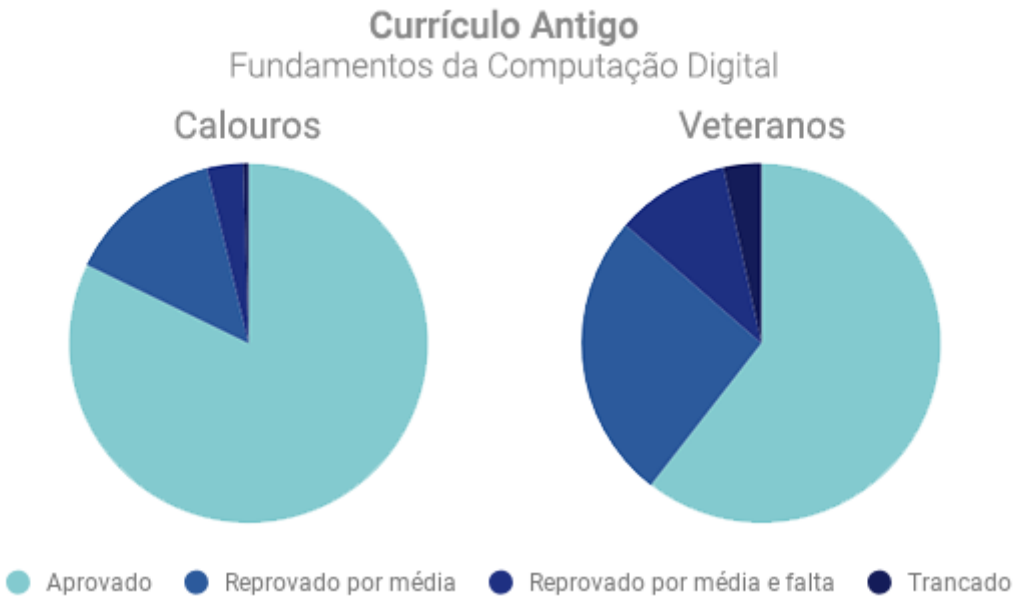


Figura 34: Currículo Antigo - Situação em Fund. da Computação Digital

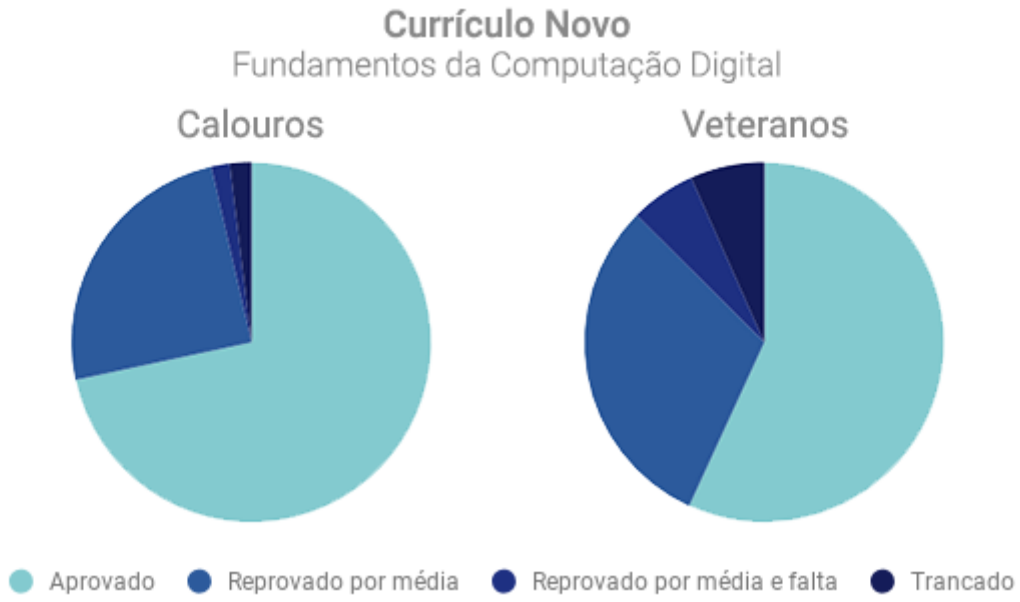


Figura 35: Currículo Novo - Situação em Fund. da Computação Digital

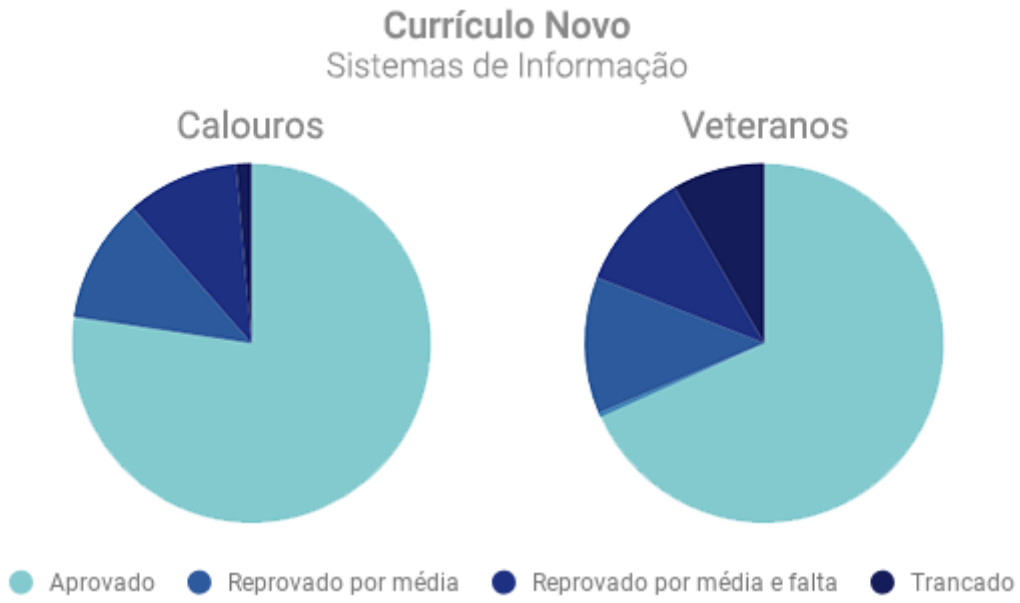


Figura 36: Currículo Novo - Situação em Sistemas de Informação